INFERENCE AND CAUSALITY

LIBF

LEARNING OBJECTIVES

UNIT 1

STATISTICAL INFERENCE

1. STATISTICAL INFERENCE

Introduction/Case study

Statistical inference allows us to make predictions about a system we want to study. In a nutshell, we assume that this system can be described by random variables that follow a specific probability density function. For example, if we want to analyze the shopping behavior of customers in a supermarket, we can, in general, assume that the demand follows a Poisson process and can be described using a negative binomial distribution. Once we have estimated the parameters of this distribution, for example, by comparing the distribution to the observed data in a fit, we can then infer the most probable or the expected demand for future sales.

In classical statistical inference we typically follow the frequentists' school of thought where we interpret the data as realizations of repeatable experiments.

This implies that we assume that the data are random, or more precisely, the data are the concrete realizations of the random variable (such as, e.g. "demand"). The parameter(s) of the probability distribution that describes the random variable, however, are fixed, even if we do not know their value. This means, in particular, that the parameters of the probability distribution are not random variables.

In Bayesian inference, on the other hand, we take a different viewpoint: Here, we assume that the data are fixed—they are what we observe and they do not necessarily have to originate from repeatable experiments. In some cases, for example, if we consider the rolling of dice, the experiment is—in principle—repeatable, if we are able to control the environment in which we perform the experiment sufficiently well. In other cases, the data are the result of single events. For example, today's weather is only observable today—we neither have access to multiple other earths with the same configuration, nor can we go back in time to observe how a hypothetical "today" might have unfolded. With the data fixed, the parameters (θ) that describe the system we want to study are now random variables. Even before we look at the recorded data, we will, in most cases, know something about the system under study. This knowledge is encoded in the prior $f(\theta)$ and may come, for example, from empirical studies performed earlier, expert knowledge, etc. Using Bayes's theorem as the core ingredient of Bayesian statistical inference, we want to determine the posterior distribution $f(\theta|x)$, that describes the probability distribution of the quantify of interest, depending on the parameters θ, given the observation of the data (x). The prior distribution $f(\theta)$ and the posterior distribution $f(\theta|x)$ can either be discrete or continuous. The parameters θ themselves will, in general, be continuous in any case.

Once we have calculated the posterior distribution, we can use this to make inferences about the system under study. For example, we can calculate the expected value of the quantity we are interested in. Coming back to the example of the supermarket, we use all the recorded sales data we have collected in the past, choose a suitable prior to calculate the posterior distribution that describes the probability density function for the future

sales in the supermarket. Using this distribution, we can, for example, calculate the expected value (or any other quantile) to estimate how many items need to be ordered to be able to fulfill the future demand.

1.1 Bayesian Inference

In most cases, we do not know that some event A will happen with certainty.

Instead, we use the probability $P(A)$ with $0 \leq P(A) \leq 1$ to express the notion that the event will occur with some probability where $P(A) = 0$ means that we are absolutely certain that the event will never occur and $P(A) = 1$ means that we are absolutely certain that the event will occur.

If we have two events, A and B, then they can be mutually exclusive, i.e., $P(A \vee B) = P(A) + P(B)$, which means that the probability that event A or event B occurs is given by the individual probabilities that each event occurs on its own. We can also calculate the probability that both events happen at the same time (for independent events): $P(A \wedge B) = P(A) \cdot P(B)$.

It is common for "$\wedge$" to be replaced by a comma, i.e., $P(A \wedge B) = P(A, B)$.

Quite often, it is difficult to directly determine the total probability of an event. In some cases, it might be possible (or easier) to determine the probability that some event A occurs at the same time as event B. If the events B_i are mutually exclusive and cover all possibilities, we can "partition" the event A:

$$P(A) = \sum_i P(A, B_i) \tag{1.1}$$

The above equation is also known as the "total law of probabilities."

The **conditional probability** $P(A \,|\, B)$ (read: probability of A, given B) means that B has already occurred and we know the values of any associated parameters. If A and B are independent, then $P(A \,|\, B) = P(A)$.

Conditional probability
The conditional probability is the probability of an event A given that an event B has already occurred or is assumed to be true.

Using the conditional probability, we can express the probability that both event A and B occur as $P(A, B) = P(A \,|\, B)P(B)$, and the total law of probabilities becomes

$$P(A) = \sum_i P(A, B_i)P(B_i) \tag{1.2}$$

Note that $P(A \,|\, B) \neq P(B \,|\, A)$. Instead, $P(A \,|\, B)P(B) = P(B \,|\, A)P(A)$, which leads to Bayes' theorem (Bayes, 1763):

$$P(A \,|\, B) = \frac{P(B \,|\, A)P(A)}{P(B)} \tag{1.3}$$

In many cases, $P(B)$ is difficult to obtain, and we use the total law of probabilities to partition over the events A_j:

$$P(A_j \,|\, B) = \frac{P(B \,|\, A_j)P(A_j)}{\sum_i P(B \,|\, A_i)P(A_i)} \tag{1.4}$$

An important application of Bayes' theorem is hypothesis testing, i.e., if we want to determine whether the data we observe can support a given hypothesis. In this case, we set $A = H$ (where H denotes the hypothesis we want to test) and $B = D$ (where D represents our data). Bayes' theorem then becomes

$$P(H \,|\, D) = \frac{P(D \,|\, H)P(H)}{P(D)}$$

and the elements of the theorem have the following meanings:

- $P(H)$—Prior. This is what we know about the system before we look at any data.
- $P(D)$—Evidence. This is the distribution of the data, fixed for a given data-set. Hence, it acts as a normalization.
- $P(D \,|\, H)$—Likelihood. This is the conditional probability of observing the data given the hypothesis, i.e., how likely is it to observe the data we have for a given hypothesis. This probability is maximal if we choose the correct hypothesis.
- $P(H \,|\, D)$—Posterior. This is what we really want to know. Given the data we observe, what is the (conditional) probability that the hypothesis we investigate is correct? Let's illustrate this with an example focused on medical diagnosis.

EXAMPLE: MEDICAL TEST

Suppose a person is not in a risk group for contracting a specific disease. A test exists for this disease and, if a person has the disease, the test will return a positive result with 99.9 percent probability.

The test will only be positive in 0.5 percent of cases, even if a patient does not have the disease.

Suppose we consider a disease with very severe consequences and the test is positive. Should the person worry?

As a first step, we translate the given probabilities into the language of statistics using conditional probabilities. We use the following notation: $+$ means the test is positive, $-$ means the test is negative, D means the patient (truly) has the disease; $\neg D$ means the patient does not have the disease. Keeping in mind that all

probabilities need to be normalized to one, we obtain

$$P(+\,|\,D) = 0.999 \qquad P(-\,|\,D) = 0.001$$
$$P(+\,|\,\neg D) = 0.0 \qquad P(-\,|\,\neg D) = 0.995$$

Using Bayes' theorem, we can calculate the posterior we want to know, i.e., if the test is positive, what is the probability that the patient has the disease?

$$P(D\,|\,+\,) = \frac{P(+\,|\,D)P(D)}{P(+\,|\,D)P(D) + P(+\,|\,\neg D)P(\neg D)}$$

If we examine this equation, we note the following: it is difficult to determine the denominator $P(\,+\,)$ describing the probability that the test will be positive. However, we can use the total law of probabilities to express this for the cases $P(D)$ and $P(\neg D)$.

We now find that we are missing a crucial piece of information: the value of the prior $P(D)$ that describes the occurrence of the disease in the population of interest. We have to get these details externally.

Knowing the accuracy of the test is not sufficient to determine whether or or not the patient has the disease.

We stated initially that the patient does not belong to a risk group.

In our example, we may refer to a database that lists how many cases there are, in a given population, among those not in a risk group. Suppose we find that the probability is $P(D) = 0.0001$, i.e., the probability of contracting the disease is very low if a patient does not belong to a risk group.

If we now put all numbers into Bayes' theorem, we obtain $P(D\,|\,+\,) = 0.02$, i.e., even if the test is positive, the probability that the person has the disease is only 2 percent.

1.1.1. Posterior Distribution

We have considered the case where we can group events into separate and discrete classes (e.g., a test is positive or negative, or a patient has 5 a specific disease). However, in most cases, we want to analyze a more complex system where we observe a continuum of values. In order to make any predictions about a system, we need a model that depends on one or more parameters θ. The "frequentist" approach to statistics assumes that while we may not know the value of the parameter(s) θ that describe the system we want to analyze, its value is fixed. In the Bayesian view however, we treat the parameter(s) θ as a **random variable** that follow a prior distribution $f(\theta)$. We then observe the data with specific values. If X is the variable describing the data and x is the observed value, we can write this as $X = x$, i.e., in the concrete realization, we observe the value x of the random varia-

ble X. For simplicity, we continue the case of just one variable, even though the same arguments hold for a vector of variables with corresponding observations: $\overrightarrow{X} = (X_1, X_2, X_3, ...) = \overrightarrow{x} = (x_1, x_2, x_2, ...)$

As was true concerning the several event categories, we are interested in the posterior distribution $f(\theta|x)$ (or, more generally, $f\left(\overrightarrow{\theta}\,\middle|\,\overrightarrow{x}\right)$). This means when we assume a specific prior distribution $f(\theta)$, the distribution describes the probability of observing a value θ for our model, conditional on the observation of the data x.

Following Eqn. (1.3), we can use Bayes' theorem and express the posterior distribution for a continuous parameter θ as:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \tag{1.5}$$

As before, the quantity $f(x|\theta) = L(\theta)$ is the likelihood function that describes the conditional probability of observing the data for a given choice of the parameter(s) θ. The function $f(\theta)$ is the prior that includes all our knowledge about θ before we analyze the data. The evidence $f(x)$ in the denominator is the normalization and describes the probability of observing the data.

We have seen earlier for the case of discrete events that it is often easier to express the evidence as a sum of events using the total law of probabilities and expanded the denominator accordingly in Eqn. (1.4). In the case of the continuous parameter θ, we can follow the same approach. However, the sum is now replaced by an integral over θ. Therefore, the posterior distribution is given by

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \tag{1.6}$$

Here, we have used

$$\int f(x|\theta)f(\theta)d\theta = \int f(x, \theta)d\theta = f(x) \tag{1.7}$$

for the denominator. Hence, we can say that the posterior distribution is given by the multiplication of the likelihood with the prior distribution, followed by normalization (Held, 2008, p. 140).

1.1.2. The Role of the Prior

We have already seen in the example above that the prior $P(A)$ for the discrete case, or $f(\theta)$ in the continuous case, plays a vital role in Bayesian inference. Recalling the example of a medical diagnosis above, we saw that we can only answer the of question whether the patient has the disease if we also know the prior: in this case, the prevalence of the disease in the non-risk population.

However, the question remains: generally speaking, from where can we obtain the prior? In some cases, we may have external or domain knowledge about the system we want to analyze. For example, when rolling dice, we may assume that that each die is fair. Hence, the (prior) probability that each side faces up is 1/6. In other cases, we may have historic data, census information or any other form of recorded statistics that, as in the medical example, allows us to determine the prior probability.

We have, however, seen that we will need to evaluate integrals of the form likelihood times prior both in the normalization of the posterior distribution (the evidence), as well as when we use the posterior distribution for inference. If possible, we would like to take a pragmatic approach and choose a prior distribution that makes the evaluation of these integrals easier. We cannot influence the parametrization of the likelihood much because this is defined by the system from which we obtain the data. We can, however, choose the form the prior takes and choose, for example, a parametrization such that the posterior calculated from the integral over the likelihood times the prior belongs to the same family of distributions as the prior. This has the advantage that the posterior distribution can expressed in a closed form and instead of using a numerical approximation, we can estimate the parameters of this posterior distribution and work with the analytic expression. We call these choices of priors conjugate priors.

> **CONJUGATE PRIOR**
> A class of priors is called a conjugate prior with respect to a given likelihood function, if the a posteriori distribution is of the same family of probability distributions as the prior.

The theory of conjugate priors was first developed in (Raiffa & Schlaifer, 1961). It is important to keep in mind that, ultimately, choosing a conjugate prior is a convenience - if we can describe our prior knowledge in terms of a conjugate prior, then we can make the further handling of the Bayes' formula easier. In other cases, however, it may not be possible to make such a convenient choice.

The most important conjugate priors are given below (Held, 2008, p. 148):

Likelihood	Conjugate prior	Prior hyper-parameter
Binomial, Bernoulli	Beta	α, β
Negative Binomial	Beta	α, β
Poisson	Gamma	α, β
Exponential	Gamma	α, β
Normal (σ^2 known)	Normal	μ, σ^2
Normal (μ known)	Inverse Gamma	α, β

Depending on the problem at hand, a specific choice of prior may be helpful.

For example, we can interpret A/B tests as a sequence of Bernoulli trials where the results fall in either category A or B. We can then choose a Beta prior, where we can interpret the hyper-parameters as $\alpha - 1$ successes and $\beta - 1$ failures in the observed data. As a special case, the choice of $\alpha = \beta = 1$ results in a flat or uniform beta distribution. As we add more data, we can interpret that we start from a uniform prior where we do not assume any knowledge about the outcome and then use $\alpha = 1$ + number of successes and $\beta = 1$ + number of failures (or choice of A and B) to refine our prior.

This raises the question of whether it is generally advisable to start with a uniform prior. Naïvely, this seems like an obvious choice: if we do not know anything a priori about the parameters of our model, it seems conservative or cautious to use a uniform prior to indicate that we do not know what their values should be.

Unfortunately, this is not the case. Many real-world datasets follow Benford's law (Newcomb, 1881; Benford, 1938), which states that the first digit of a number follows a skewed distribution given by Eqn. (1.8). This means that numbers start more frequently with a one than a two, three or any other digit—and the same holds for the other digits. Hence, only few numbers start with a nine. In other words, in nature the logarithms of the numbers is uniformly distributed—not the numbers themselves.

$$P(d) = \log_{10}\left(1 + \tfrac{1}{d}\right) \tag{1.8}$$

Instead, we want the prior we choose to contain as little information as possible. Apart from Benford's law, the uniform distribution is not ideal for this. Consider, for example, the case that we start with a uniform distribution and then choose another set of parameters or coordinates to re-parametrize the distribution. Since we only change the way we express the parameters (but not what they represent), we expect that this has no consequence on the prior. Let $\phi = h(\theta)$ be the transformation where the function h transforms the original parameters θ to a new parametrization ϕ. If the function h is bijective, we have a 1:1 correspondence between ϕ and θ. The distribution of ϕ is then given by Eqn. (1.9) (Held, 2008, p. 151). However, unless the transformation h is linear, the resulting distribution $f(\phi)$ is not constant, even if we start with a uniform distribution for θ. Hence, a simple change in the way we express the parametrization transforms the uniform distribution into a different shape. This means that trying to use the uniform distribution to express that we do not know much about the parameters is not helpful.

$$f(\phi) = f\left(h^{-1}(\phi)\right)\left|\frac{dh^{-1}(\phi)}{d\phi}\right| \tag{1.9}$$

Instead, we look for a non-informative prior such as the Jeffrey prior (Jeffreys, 1946), see also (Liu & Abeyratne, 2019, App. 4) or (Gelman & Rubin, 1992, p. 52ff). This is defined as

$$f(\theta) \propto \sqrt{J(\theta)} \tag{1.10}$$

where $J(\theta)$ is the expected **Fisher information** of θ. The Fisher information measures the amount of information about the parameters θ and is given by the negative of the second derivative of the log-likelihood function:

$$I(\theta) = -\frac{d^2\text{LogL}(\theta)}{d\theta^2} \tag{1.11}$$

The first derivative of the log-likelihood function is also called the "score function" $S(\theta)$:

$$S(\theta) = \frac{d\text{LogL}(\theta)}{d\theta} \tag{1.12}$$

The Fisher information can then be written as

$$I(\theta) = -\frac{d^2\text{LogL}(\theta)}{d\theta^2} = -\frac{dS(\theta)}{d\theta} \tag{1.13}$$

The expected Fisher information is then the expectation value of $I(\theta)$, i.e.,

$$J(\theta) = E[I(\theta)] \tag{1.14}$$

Under the assumption that we can change the order of differentiation and integration (regularization assumption), we can show that (Held, 2008, p. 66):

$$E[S(\theta)] = 0 \tag{1.15}$$

$$Var[S(\theta)] = E[S(\theta)^2] = J(\theta) \tag{1.16}$$

Using the transformation rule in Eqn. (1.9), we can show that the Jeffrey prior has the same form before and after the transformation:

📖 **THE JEFFREY PRIOR IS INVARIANT UNDER BIJECTIVE TRANSFORMATIONS**

Show that the Jeffrey prior is invariant under bijective transformations.

We define the Jeffrey prior for the parameter θ as $f(\theta) \propto \sqrt{J(\theta)}$ according to Eqn. (1.10). Then, we use the rule for the transformation of probability distributions in Eqn. (1.9):

Fisher information
The Fisher information is a metric that measures the amount of information that a random variable X contains.

$$f(\phi) \propto f\left(h^{-1}(\phi)\right)\left|\frac{dh^{-1}(\phi)}{d\phi}\right|$$

$$\propto f(\theta)\left|\frac{dh^{-1}(\phi)}{d\phi}\right| \quad \text{with} \quad f(\theta) = f\left(h^{-1}(\phi)\right)$$

$$\propto \sqrt{J(\theta)}\left|\frac{dh^{-1}(\phi)}{d\phi}\right| \quad \text{with} \quad f(\theta) \propto \sqrt{J(\theta)}$$

$$= \sqrt{J(\theta)\left|\frac{dh^{-1}(\phi)}{d\phi}\right|^2}$$

$$= \sqrt{J(\phi)}$$

Hence, if we express the prior $f(\theta)$ according to Jeffrey's rule and then transform $\theta \to \phi$, the resulting prior using the transformed variable also follows Jeffrey's rule (Held, 2008, p. 152).

This allows us to construct a prior that does not depend on the parametrization chosen for the distribution of the parameter(s) θ that describes our model of the system we wish to analyze.

1.1.3. Bayesian Prediction

Once we have determined the posterior distribution, we need to derive quantities that we can use, for example, as a concrete prediction: The full posterior distribution includes all knowledge we have of the system we want to study, including the expected volatility or uncertainty. However, in many practical scenarios, we need a point estimate. In principle, we can use any quantile of the distribution, however, the following point estimators are most commonly used:

- The expectation value is given by

$$E[\theta \,|\, x] = \int \theta f(\theta \,|\, x) d\theta \tag{1.17}$$

- The mode is the maximum of the a posteriori distribution

$$Mod(\theta \,|\, x) = \operatorname{argmax}_\theta f(\theta \,|\, x) \tag{1.18}$$

- The median is the quantile that cuts the posterior distribution in half, i.e., 50 percent of the distribution are on the left of this quantile, 50 percent on the right:

$$\int_{-\infty}^{q_{0.5}} f(\theta \,|\, x) d\theta = 0.5 \quad \text{and} \quad \int_{q_{0.5}}^{\infty} f(\theta \,|\, x) d\theta = 0.5 \tag{1.19}$$

The best choice of the point estimator depends on the problem at hand.

As usual, the mode is quite sensitive to the exact shape of the estimated posterior distribution, and small fluctuations may have a big impact. The expectation value may be influenced by long tails of the a *posteriori* distribution, whereas the median is generally more stable.

Additionally, we can construct credible intervals, i.e., regions that contain the variable $\theta \,|\, x$ with probability $1 - \alpha$. These intervals are defined by an lower and an upper bound.

$$\int_{b_l}^{b_u} f(\theta \,|\, x) d\theta = 1 - \alpha \tag{1.20}$$

The easiest way to set the boundaries is to use the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution.

Note that the credible interval is similar to (but not the same as) the confidence interval used in frequentist statistics. Remember that in frequentist statistics, the parameter(s) θ are unknown and fixed, whereas in Bayesian inference, the parameter(s) θ are random variables. In the frequentist's view we say that if we repeat an experiment many times, the unknown parameter θ will be contained in the confidence interval in $100 \cdot (1 - \alpha)\,\%$ of the cases. For example, a 95 percent confidence interval means that if we repeat the experiment very often, the (fixed) parameter θ would be in that interval in 95 percent of cases. We cannot make a claim that the parameter θ is contained within the confidence interval with probability $1 - \alpha$, because this is a Bayesian interpretation and not defined within the frequentist's interpretation.

In the Bayesian credible interval ~~on the other hand~~, we have access to the posterior distribution directly and can hence assert that the parameter will be in the region bounded by b_l and b_u with 95 percent probability.

Note, however, that the Bayesian credible intervals contain additional information via the prior distribution, whereas the confidence intervals are constructed from data alone.

1.2 Bayesian Networks

So far, we have encountered systems where we need maybe one or two pieces of information to infer the probability of an event (e.g., if a patient contracts a disease despite not belonging to a risk group and their test for that disease is positive). In many situations, we need to take a large number of variables into account, meaning that the probability depends on a set of variables $P(X_1, X_2, ..., X_n)$. Even if each variable is binary and can be expressed by zero or one (or true or false), we would need to store $2n - 1$ elements. Apart

from practical considerations of handling these values, it is also very difficult to calculate such a joint probability that depends on many variables. To simplify this joint probability, we look for ways to split it into smaller components that we can more easily treat.

In the case of two variables X_1 and X_2, we can express the joint probability $P(X_1, X_2)$ as $P(X_1, X_2) = P(X_2 | X_1)P(X_1)$. We can generalize this with regard to more variables and obtain the chain rule for probability:

$$P(X_1, X_2, \ldots, X_n) = \prod_j P(X_j | X_1, \ldots, X_{j-1}) \tag{1.21}$$

> 📖 **EXAMPLE OF CHAIN RULE OF PROBABILITY**
> For the case of four variables, we can express the joint probability $P(X_1, X_2, X_3, X_4)$ as
>
> $$\begin{aligned} P(X_1, X_2, X_3, X_4) &= P(X_4 | X_3, X_2, X_1)P(X_3, X_2, X_1) \\ &= P(X_4 | X_3, X_2, X_1)P(X_3 | X_2, X_1)P(X_2, X_1) \\ &= P(X_4 | X_3, X_2, X_1)P(X_3 | X_2, X_1)P(X_2 | X_1)P(X_1) \end{aligned}$$

This by itself does not help us much, as we simply expressed the joint probability as a (long) product of conditional probabilities. However, when we build a model, we know more about the characteristics of the system we wish to describe.

Independent
Variables are independent if $P(X_1, X_2)$ $= P(X_1)P(X_2)$.

Some variables may be **independent**, which means that we have $P(X_1 | X_2) = P(X_1)$ for the conditional probability. This expresses that the probability of observing X_1 is independent of X_2. We can call this unconditional or absolute independence of the two variables. In contrast, two variables can also be conditionally independent on a third variable. We will cover conditional independence formally a bit later. For now, we say that the two variables X_1 and X_2 are conditionally independent given X_3 if $P(X_1 | X_2, X_3) = P(X_1 | X_3)$ and $P(X_2 | X_1, X_3) = P(X_2 | X_3)$. This means that if we

know the value of X_3, the variables X_1 and X_2 become independent.

Hence, if we know more about the structure of the system we wish to model, we can considerably simplify the (conditional) probabilities from the chain rule. For example, if we knew that X_1 became conditionally independent of all other variables based on knowing the value of X_2, we can write

$$P(X_1, X_2, ..., X_n) = P(X_1 | X_2)P(X_2, ... X_n)$$

instead of

$$P(X_1, X_2, ..., X_n) = P(X_1 | X_2)P(X_2, ...X_n)$$

Using our expert or domain knowledge, we can make the relations between variables explicit. Suppose we have three variables, A, B, and C, and we know that both A and C depend on B, but A does not depend on C.

Hence, we can say

- A is conditionally dependent on B: $P(A | B)$.
- C is conditionally dependent on B: $P(C | B)$.
- A is conditionally independent from C given B: $P(A | C, B) = P(A | B)$.
- C is conditionally independent from A given B: $P(C | A, B) = P(C | B)$.

Therefore, we can express the joint probability as

$$
\begin{aligned}
P(A, B, C) &= P(A | B, C)P(C | B)P(B) \quad \text{chain rule} \\
&= P(A | B)P(C | B)P(B) \quad \text{cond. independence}
\end{aligned}
$$

We can visualize these relations as shown in figure below; this is a simple Bayesian network. Bayesian networks are graphical representations of the statistical relations between variables and were introduced by Pearl in the 1980s (Pearl, 1985; Pearl & Russel, 2003; Pearl, 2014a). Technically, a Bayesian network is represented as a directed acyclic graph. For now, we say that the nodes (shown as boxes) represent variables that are connected by "edges" (shown as arrows), indicating the relationship between variables.

Figure 1: A Simple Bayesian Network
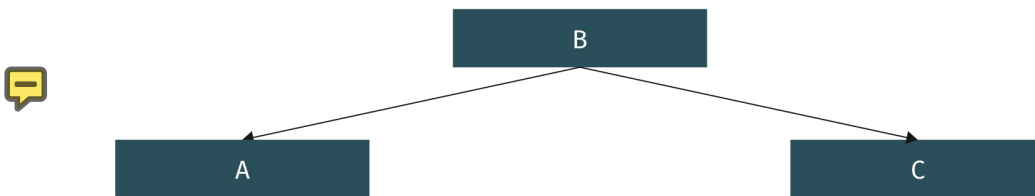
In this graph, B is the top node and is called the "parent" of both A and C.

Parent variables are often denoted with as PA, i.e., the parent of variable X_j for some index j is PA_j. A formal definition is given in (Pearl, 2009, p. 14).

Note that we do not require the connections to represent causal relationships: although using our domain knowledge we find that many of these relationships have a causal meaning, we can express non-causal structures in Bayesian networks. They are meant as a method to simplify working with the joint probability, regardless of whether a causal relationship exists between variables.

So far, we have used the Bayesian networks to represent the relationship between variables. In order to infer the values of variables or determine the probability of outcomes, we need to add the concrete values for all (conditional) dependencies. These tabulated values are called conditional probability tables (CPT), which summarize the values we observe in the data. In the simplest case, all variables are binary and can be represented in terms of "true" or "false". figure below shows a simple example that has been adapted from (Murphy, 2001). In this example, we want to express the probability that the grass is wet, which is related to either a sprinkler or rain. In this network, the sky condition is the parent of all nodes. It can be cloudy (or not) each with a probability of 50 percent. The variables "sprinkler" and "rain" are then children of "cloudy." Their values depend on the value of "cloudy." Hence, each value of the variable (e.g. "sprinkler = true" or "sprinkler = false") depends on the value of the parent, (i.e., "cloudy = true" or "cloudy=false"). The same applies for the variable "rain." The final variable "wet grass" can again take the value "true" or "false"—but now this value depends on the value of both its parents "sprinkler" and "rain." Hence, the conditional probability table for "wet grass" needs to capture all combinations where each variable is either "true" or "false." Note that the numbers in the CPT for the wet grass example are fictional values. In the case of a real system, we would carefully measure all conditional probabilities that enter the modeling of the system. A more detailed example is shown in figure below, which shows the "Asia network" (Lauritzen & Spiegelhalter, 1988).

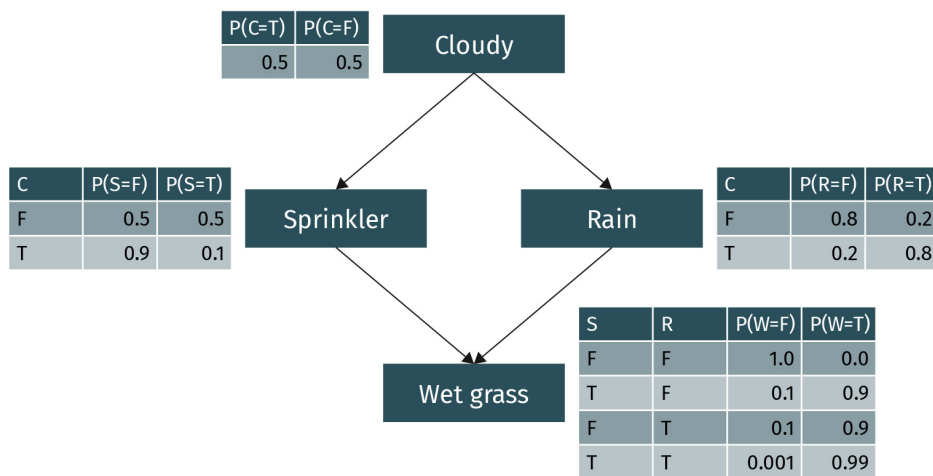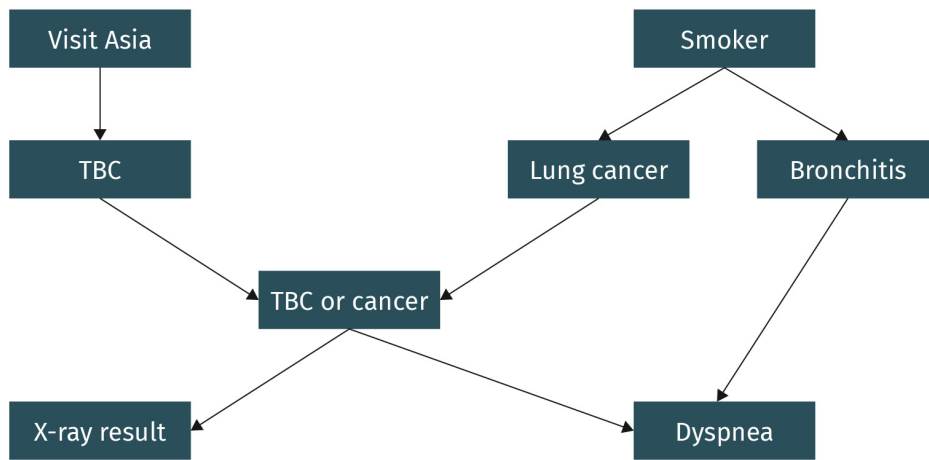**Figure 2: Bayes Network for Wet Grass, adapted from (Murphy, 2001)**

Figure 3: Asia Network, adapted from (Lauritzen & Spiegelhalter, 1988)

This is used to determine the probability of the quantity of interest. On its own, each part of this Bayesian network contains the conditional probabilities of visiting Asia. Say that in 99 percent of the cases, a person has not visited Asia. Then, following all conditional probabilities, we can determine the probability that a person has an abnormal X-ray result or suffers from dyspnea. We can also turn this around and determine how likely it is that a person with dyspnea also smokes. Since the arrows do not represent causal relationships, we can move across the graph to investigate the conditional probabilities as they change depending on the values of other variables. For example, if we know that a person has dyspnea, we can set the value of this variable to 100% and then use the conditional dependencies to observe the change in all other variables.

This allows us to use Bayesian networks to reason under uncertainties.

As we model all dependencies in conditional probability tables, we can determine the effect of a given or imagined observation. Using the example of the wet grass, we can also ask whether it is more likely that the grass is wet because the sprinkler has been switched on or because it has been raining. Hence, we want to determine the conditional probabilities $P(R = T | W = T)$, i.e., what is the probability that it has been raining, given that we see that the grass is wet, as well as the $P(S = T | W = T)$ for the sprinkler. Using the chain rule, the conditional probability for the sprinkler having gone off is given by

$$P(S = T | W = T) = \frac{P(S = T, W = T)}{P(W = T)} \qquad (1.22)$$

meaning that the conditional probability is the joint probability divided by a normalization factor expressing that we observe the grass to be wet ($P(W = T)$). To calculate this quantity, we need to sum over all possible values (or integrate in case of continuous variables) of the joint probability.

This process is called marginalization, i.e., we wish to obtain the **marginal distribution**. In wet grass example, we need to sum over all possible values (i.e., "true" and "false") of all variables except the one describing the wet grass. This variable is set to one, as we observe the grass to be wet.

We use the notation where capital letters (C, S, R) refer to the variables and small letters (c, s, r) indicate the values of the variables, in our case "true" and "false." We need to compute the expression:

$$P(W = T) = \sum_c \sum_s \sum_r P(C = c, S = s, R = r, W = T) \tag{1.23}$$

To allow use of the conditional probability tables in the Bayesian network, we need to first use the chain rule to expand the joint probability into conditional probabilities:

$$P(C, S, R, W) = P(C) \cdot P(S|C) \cdot P(R|C, S) \cdot P(W|C, S, R) \tag{1.24}$$

From our model representing the graph, we know that the sprinkler S and rain R are conditionally independent, given the clouds C, Hence we can simplify $P(R|C, S) = P(R|C)$. We also know that the value of C is no longer necessary once we know the state of S and R, meaning that we can write $P(W|C, S, R) = P(W|S, R)$. Now, the expansion for the joint probability becomes

$$P(C, S, R, W) = P(C) \cdot P(S|C) \cdot P(R|C) \cdot P(W|S, R) \tag{1.25}$$

which are the values we have in our conditional probability tables. Hence the equation for $P(W = T)$ becomes

$$P(W = T)$$

$$= \sum_c \sum_s \sum_r P(C = c, S = s, R = r, W = T)$$

$$= sum_c \sum_s \sum_r P(C) \cdot P(S|C) \cdot P(R|C) \cdot P(W = T|S, R)$$

and we can calculate this explicitly:

$$P(W = T)$$
$$= P(C = F) \cdot P(S = F \mid C = F) \cdot P(R = F \mid C = F) \cdot P(W = T \mid S = F, R = F) +$$
$$P(C = F) \cdot P(S = F \mid C = F) \cdot P(R = T \mid C = F) \cdot P(W = T \mid S = F, R = T) +$$
$$P(C = F) \cdot P(S = T \mid C = F) \cdot P(R = F \mid C = F) \cdot P(W = T \mid S = T, R = F) +$$
$$P(C = F) \cdot P(S = T \mid C = F) \cdot P(R = T \mid C = F) \cdot P(W = T \mid S = T, R = T) +$$
$$P(C = T) \cdot P(S = F \mid C = T) \cdot P(R = F \mid C = T) \cdot P(W = T \mid S = F, R = F) +$$
$$P(C = T) \cdot P(S = F \mid C = T) \cdot P(R = T \mid C = T) \cdot P(W = T \mid S = F, R = T) +$$
$$P(C = T) \cdot P(S = T \mid C = T) \cdot P(R = F \mid C = T) \cdot P(W = T \mid S = T, R = F) +$$
$$P(C = T) \cdot P(S = T \mid C = T) \cdot P(R = T \mid C = T) \cdot P(W = T \mid S = T, R = T)$$
$$= 0.5 \cdot 0.5 \cdot 0.8 \cdot 0.0 +$$
$$0.5 \cdot 0.5 \cdot 0.2 \cdot 0.9 +$$
$$0.5 \cdot 0.5 \cdot 0.8 \cdot 0.9 +$$
$$0.5 \cdot 0.5 \cdot 0.2 \cdot 0.99 +$$
$$0.5 \cdot 0.9 \cdot 0.2 \cdot 0.0 +$$
$$0.5 \cdot 0.9 \cdot 0.8 \cdot 0.9 +$$
$$0.5 \cdot 0.1 \cdot 0.2 \cdot 0.9 +$$
$$0.5 \cdot 0.1 \cdot 0.8 \cdot 0.99 +$$
$$= 0 \ + \ 0.045 \ + \ 0.18 \ + \ 0.0495 \ +$$
$$0 \ + \ 0.324 \ + \ 0.009 \ + \ 0.0396$$
$$= 0.6471$$

If we then want to calculate the probability that the sprinkler was on ($S = T$) after we observe the wet grass ($W = T$), we need to calculate $P(S = T \mid W = T)$ as given in Eqn. (1.22). Hence, we need to calculate the joint probability $P(S = T, W = T)$ in the same way as we have obtained the normalization constant above. This time, we only need to sum over c and r, since we set $W = T$ and $S = T$: $P(S = T, W = T) = \sum_c \sum_r P(C = c, S = T, R = r, W = T)$ and, following the same approach as above, we obtain $P(S = T, W = T) = 0.2781$. Therefore, the probability $P(S = T \mid W = T) = 0.2781 / 0.6471 = 0.430$.

In a more complex graph (like the one concerning the Asia network), we can then explore the effect of setting various variables to different values and observe how the other variables change. This allows us to calculate the probability of the effect we wish to investigate given all the other variables.

Remember that we do not make any assumptions about causal relationships at this point, although the conditional probabilities will often reflect a causal structure.

1.3 Probabilistic Modelling

When discussing Bayesian modeling earlier, we saw in Eqn. (1.6) that we need to estimate the posterior distribution $f(\theta\,|\,x)$ to make inferences about the system we are interested in, where the posterior is given by

$$f(\theta\,|\,x) = \frac{f(x\,|\,\theta)f(\theta)}{\int f(x\,|\,\theta)f(\theta)d\theta}$$

For any concrete prediction, we need to choose a suitable point estimator.

We specify this estimator using a loss function (sometimes also called cost function) $C(a,\theta)$ that specifies the loss (or cost) if we estimate θ using a.

Then the Bayes estimator of θ with respect to this cost function is given by:

$$E[C(a,\theta)\,|\,x] = \int C(a,\theta)f(\theta\,|\,x)d\theta$$

In the simplest case, we use the expectation value of the posterior distribution:

$$E[\theta\,|\,x] = \int \theta f(\theta\,|\,x)d\theta$$

Conjugate priors
The conjugate prior is from the same family of distributions as the posterior.

As discussed earlier, in some cases we can choose a suitable prior distribution such that the posterior distribution is known - this leads us to the concept of **conjugate priors**.

However, this only works in select cases and, in general, we need to calculate the integral. Unfortunately, this can only be done analytically for a limited number of functions or when the integrand can be transformed such that an analytical solution is known. In many cases, the integral has to be evaluated numerically. This can be done using random numbers, which is why this method is also called "Monte-Carlo integration," where "Monte-Carlo" refers to the famous casinos in Monte-Carlo, a hint to the random numbers used in the process.

First, we investigate the one-dimensional case where we have to evaluate integrals of the form

$$I = \int_a^b f(x)dx$$

which can be brought to the standardized form

$$I = \int_0^1 f(x)dx$$

PREVIEW-PDF, erzeugt: 2024-06-18T14:30:01.73+02:00

using a suitable variable transformation. The simplest approach is then to interpret the integral as the constant function 1 and determine the expectation value

$$I = \int_0^1 1 \cdot f(x)dx = E[f(x)]$$

We then choose random numbers and we evaluate the function $f(x)$ at these specific values. Then, we approximate the expectation value with the sample mean:

$$I = E[f(x)] \approx \frac{1}{n}\sum_i f(x_i) \tag{1.26}$$

The more random numbers we use in this procedure, the more accurate the estimate will be. Unfortunately, in many cases obtaining a sample from the function $f(x)$ may be difficult to obtain. For example, we may not be able to sample the function directly because we do not know the complete parametrization of $f(x)$ or it may be difficult to do so. When evaluating the sum in Eqn. (1.26) above, we implicitly assumed that we choose the values x_i at which we evaluate the function, from a uniform distribution. However, if the function $f(x)$ varies rapidly or is concentrated in a small region, this is not very efficient, since a large number of samples do not contribute much to the final result. It would be better to choose our sampling points such that more samples are drawn from a region where $f(x)$ is concentrated in. This challenge is amplified in higher dimensions. In these cases we can evaluate the integral using a technique called importance sampling.

The intuition behind this approach is to find a suitable function $g(x)$ from which we can sample. The function $g(x)$ should be defined on the same interval as $f(x)$ and mimic $f(x)$ as closely as possible while being easier to evaluate.

Then, we can use

$$I = \int_0^1 f(x)dx = \int_0^1 \frac{f(x)}{g(x)}g(x)dx = E\left[\frac{f(x)}{g(x)}\right]$$

and, again approximating the expectation value with the sample mean, we obtain

$$I \approx \frac{1}{n}\sum_i \frac{f(x_i)}{g(x_i)} \tag{1.27}$$

where the random numbers are chosen according to $g(x)$. Since $g(x)$ mimics the shape of $f(x)$, the distribution of random numbers we use to sample $f(x)$ more closely follows the regions where $f(x)$ changes rapidly (as compared to choosing the random numbers, e.g., according to a uniform distribution). This allows us to evaluate $f(x)$ more accurately with a lower number of random numbers.

The crude approach becomes more challenging in higher dimensions. This can be illustrated using a two-dimensional example where we have to integrate a function in the triangle given by, say, the points $(0,0), (1,0), (1,1)$:

$$I = \int_0^1 \int_0^x f(x,y) dy dx$$

Following the one-dimensional approach, we might be tempted to evaluate the integral using random numbers obtained using the approach outlined below:

- Generate a random number x_i from a uniform distribution in $(0,1)$.
- Generate a random number y_i from a uniform distribution in $((0, x_i))$.

However, as we can see in figure below, this approach does not work: the random numbers are not distributed evenly, although we have implicitly assumed this in the simple approach in order to generate the random numbers above. In particular, the numbers towards the origin $(0,0)$ are much denser than in the region towards $x = 1$. This can lead to severe biases in the evaluation of the integral. This issue becomes even more problematic in higher dimensions, and we have to use other methods to approximate the posterior distribution.

Figure 4: Random Numbers in the Triangle (0, 0), (1, 0), (1, 1)
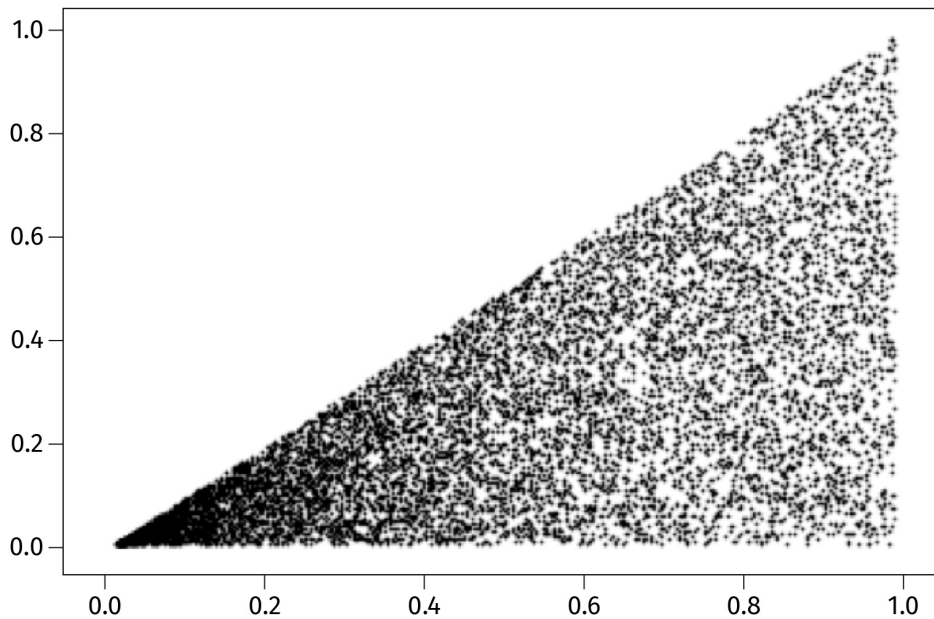
The main challenge here is that we do not generally fully know the details of the posterior distribution. Hence, we lack the means to generate statistically independent samples from this distribution. In particular, recall that the posterior distribution is given by

$$f(\theta\,|\,x) = \frac{f(x\,|\,\theta)f(\theta)}{\int f(x\,|\,\theta)f(\theta)d\theta}$$

The denominator is called the "evidence" ($P(D)$), essentially the normalization given the observed data. We need to solve at least this integral to be able to work with the posterior distribution. Hence, we need to estimate the properties of the posterior distribution without knowing everything about it. We need to know how to calculate the **likelihood**. A popular way of achieving this is called "Markov Chain Monte Carlo (MCMC) sampling" where we use random numbers ("Monte Carlo") in a special way ("Markov Chain") to draw random numbers according to the posterior distribution we are interested in. In other words, we use a special process (the Markov chains) to approximate the Monte Carlo integration we have discussed so far. To develop an intuition as to how Markov Chain Monte Carlo sampling works, we need to understand some crucial properties of Markov Chains.

Likelihood
The likelihood without knowing everything about it. We need to know how to calculate measures how well a distribution with its parameters fits the observed data.

Markov Chains are named after A.A. Markov, see e.g. (Hayes, 2013) for a historic account. Markov Chains are used to describe systems with a specific number of states. A simple example shown in figure below shows a system with three states.

Figure 5: A Simple Markov Chain.
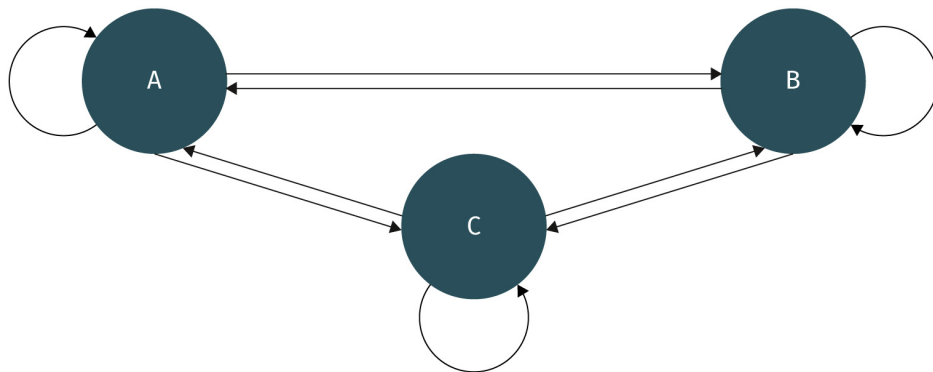
Each state is connected to another state and even to itself via an arrow that represents the probability that the state will go from state s_i to the next state s_{i+1} according to a specific probability.

For example, we can say that states A, B, C represent the weather, e.g. cloudy, rainy, or sunny. Or they might represent student life: studying, sleeping, eating, or others. Let's take the example of the weather: If we model tomorrow's weather observing today's weather, we can say that each change in weather occurs with a specific probability. For example, we may find that, if it's sunny today, there's a 60 percent probability that it will be sunny tomorrow as well, a 25 percent probability that it's cloudy tomorrow, and a 15 percent probability that it will rain. As we go from day to day, we can predict tomorrow's weather by judging today's weather. To express this in more general terms, we traverse the Markov Chain and move from state s_i to state s_{i+1}. Such a sequence might be: $A, B, B, A, C, A, C, B, B, A, A, ...$.

Memoryless
Memoryless means that the next state only depends on the current state but not all the states that precedes it.

The crucial property of Markov Chains is that they are **memoryless**. In the weather example, this means that we can make a prediction about tomorrow's weather without knowing the weather of all days preceding today. Once we are in a particular state (say, A), we can calculate the probability that we will observe any other state without knowing how we got into the current one. If we traverse the Markov Chain long enough, we will eventually reach the equilibrium or stationary state where we can predict which state we are going to be in with a given probability, regardless of the initial state. In the example of the three states, we find that, in the equilibrium case, we are in state A with p_A, in state B with p_B

and C with p_C. For example, imagine the following matrix that determines the transition from any state to the next:

$$P(s_{i+1} | s_i) = \begin{array}{c|ccc} & A & B & C \\ \hline A & 0.8 & 0.1 & 0.1 \\ B & 0.2 & 0.7 & 0.1 \\ C & 0.15 & 0.25 & 0.6 \end{array} \tag{1.28}$$

PREVIEW-PDF, erzeugt: 2024-06-18T14:30:01.73+02:00

Hence, we go from $A \to A$ with 80%, from $A \to B$ with 10% probability, etc. Note that each row adds up to one, as we need to end up in one of the states A, B or C.

The equilibrium state of this Markov Chain with the transition probabilities given by Eqn. (1.28) is $A = 0.475$, $B = 0.325$, $C = 0.2$. Figure below shows how the Markov Chain converges towards this equilibrium state, initially starting from state B. Since the Markov Chain is memoryless, the equilibrium state does not depend on which state we start from.

Figure 6: Equilibrium State of a Simple Markov Chain with Three States.
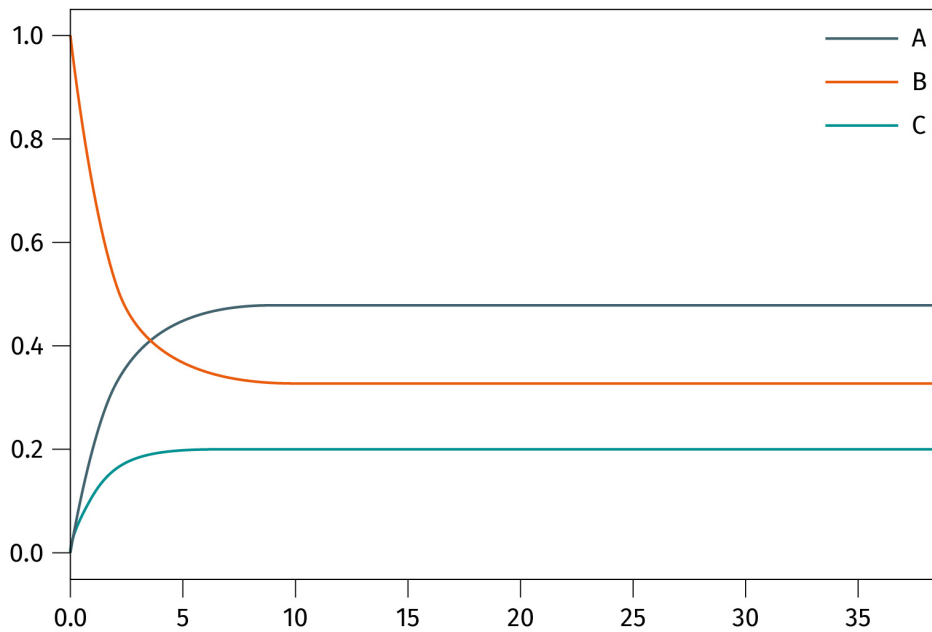
The general idea behind Markov Chain Monte Carlo is that we construct a chain that has the desired distribution (in our case, the posterior distribution), as its stationary or equilibrium point. Once we reach the equilibrium point we can then use the Markov Chain to sample from the distribution, i.e., to generate random numbers according to the shape of the posterior distribution. As we have seen above, the starting point does not matter, as we will reach the equilibrium point after the method has been applied for sufficient time. The challenge is then to find a set of states $s = (s_1, s_2, ... s_m)$ that has the distribution we are interested in as its stationary distribution, i.e., $s = Ps$, where s is the vector of states and P is the transition probability matrix.

One way to do this is the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970). Intuitively, the algorithm works as follows: Suppose we want to sample from some distribution $p(x) = \widetilde{p}(x)/C$, where $\widetilde{p}(x)$ is a distribution that is proportional to the distribution $p(x)$ we are interested in. In our case, $p(x)$ is the posterior distribution $f(\theta \,|\, x)$ and $\widetilde{p}$ is proportional to this. We then start with an arbitrary state s_x. In case of continuous distributions, this is a point (x_t) chosen randomly. Then, we repeat the following two steps:

1. Generate a new point y (or state s_y) for some Markov transition matrix $Q(y|x)$
2. Accept the point $x_{t+1} = y$ with probability

$$\alpha\{y|x\} = min\left\{1, \frac{\widetilde{p}(y)}{\widetilde{p}(x)} \cdot \frac{Q(x|y)}{Q(y|x)}\right\}$$

Otherwise, keep the old point, i.e. $x_{t+1} = x_t$

The function $\alpha(y|x)$ is called the acceptance probability and intuitively describes if the proposed new state is in a region were the desired target distribution is not vanishingly small. Then, once the algorithm converges, we can sample from the distribution $p(x)$ (in our case, the posterior distribution).

However, the above outline requires some further discussion.

First, how do we judge whether the algorithm has converged? Indeed, this is difficult to establish and, generally, needs to be assessed on a case-bycase basis. Typically, we run the algorithm for a given "warm-up" period, during which the resulting values are not recorded to get to the region where the states of the Markov Chain built by the algorithm are more representative of the distribution we want to sample from. This is discussed further in (Gelman, 2014, chap. 11.4).

The other question concerns which distribution we should choose for Q? The choice of this proposal distribution has a big impact on how long we need until we reach the equilibrium or stationary point. The exact choice also depends on the problem at hand. For the Metropolis (not the Metropolis-Hastings) algorithm, a symmetric distribution is chosen for which $Q(y|x) = Q(x|y)$. In this case, the acceptance criterion is simpler and becomes

$$\alpha(y|x) = min\left\{1, \frac{\widetilde{p}(y)}{\widetilde{p}(x)}\right\} \tag{1.29}$$

Empirically, a Gaussian or Normal distribution is often chosen with mean $\mu = x$, i.e., $Q \sim \mathcal{N}(x, \sigma^2)$. The variance $\sigma 2$ is then a parameter that has to be tuned during the warm-up period of the algorithm. This leads to a random walk where proposed points x_{t+1} around the previous point x_t are more likely.

In the acceptance criterion above, we have thus far referred to a general distribution $p(x)$ from which we would like to sample and a distribution $\widetilde{\ }p(x)$ that we use for accepting a new point where $\widetilde{p}(x) \propto p(x)$. In our case of Bayesian inference, we want to sample from the posterior distribution $f(\theta|x)$. Therefore, we need to construct a distribution that is proportional to this. Remembering Bayes' theorem (Eqn. (1.6)), we know that the posterior is proportional to the likelihood times the prior. Hence, we can express the ratio of posteriors as

$$\frac{f(\theta^*|x)}{f(\theta|x)} = \frac{\frac{f(x|\theta^*)f(\theta^*)}{\int f(x|\theta^*)f(\theta^*)d\theta^*}}{\frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}}$$

$$= \frac{f(x|\theta^*)f(\theta^*)}{f(x|\theta)f(\theta)}$$

where θ^* represents the proposed new values of the posterior distribution $f(\theta|x)$ (given

the observed data x) and θ the current values. Hence, by using Bayes' theorem we can express the ratio of posteriors we are interested in as the ratio of the likelihoods times the priors—and both quantities are known or described in our model. In particular, we notice that the denominator in Bayes' theorem representing the evidence (or normalization) drops out. This is good as we cannot in general compute it.

The Metropolis and Metropolis-Hastings algorithms are conventionally relatively easy to understand. However, in many real-world scenarios, they are not powerful enough. One of the problems is that the Metropolis- Hastings algorithm is a bit too random and, for example, has a high reject (or low acceptance) rate. This, in turn, means that a lot of computations are wasted. Other MCMC sampling techniques have to be used such as Gibbs sampling (Geman & Geman, 1984) or the more performant No-UTurn- Sampler (NUTS) (Hoffman & Gelman, 2014) using a slightly different approach called "Hamiltonian Monte Carlo" (Duane, Kennedy, Pendleton, & Roweth, 1987; Betancourt, 2017).

Further details on Markov Chains and MCMC can also be found in, for example, (Gelman, 2014; Van Ravenzwaaij, Cassey, & Brown, 2018).
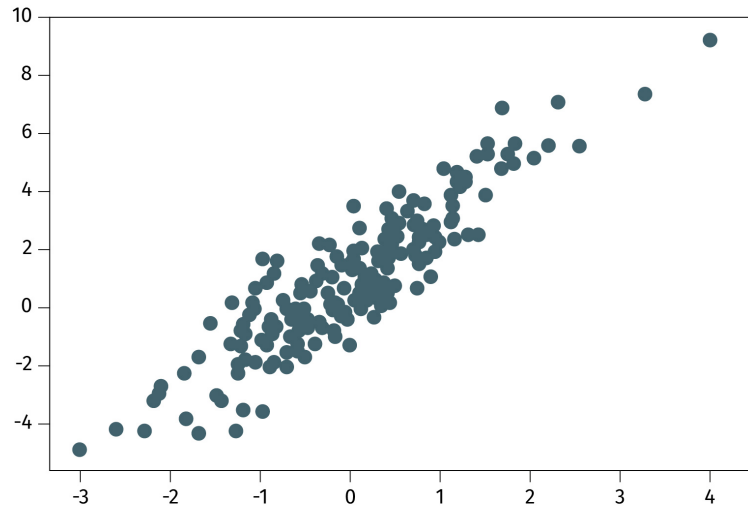
EXAMPLE: LINEAR REGRESSION

Let us now illustrate probabilistic modeling using linear regression as an example and generate some simple toy data according to the following model for n data points:

1. Generate n x-values according to a standard Normal (or Gaussian) distribution.
2. For each x-value, compute the corresponding y-value as
   $y_i = m \cdot x_i + b + \epsilon_i$ where m is the slope of a linear model, b the intercept and ϵ_i a noise term that is distributed according to a standard Normal distribution.

For $m = 2.0, b = 1.0$, the following dataset is obtained if we generate $n = 200$ samples:

In a standard linear regression approach, we would use the data and determine the model parameters, for example, via least-squared optimization.

The following plot shows the result of such a fit, together with the data and the model we used to generate the toy data. The optimization returned the values $m = 2.100$ and $b = 1.046$, which is reasonably close to the values we used to generate the dataset with.
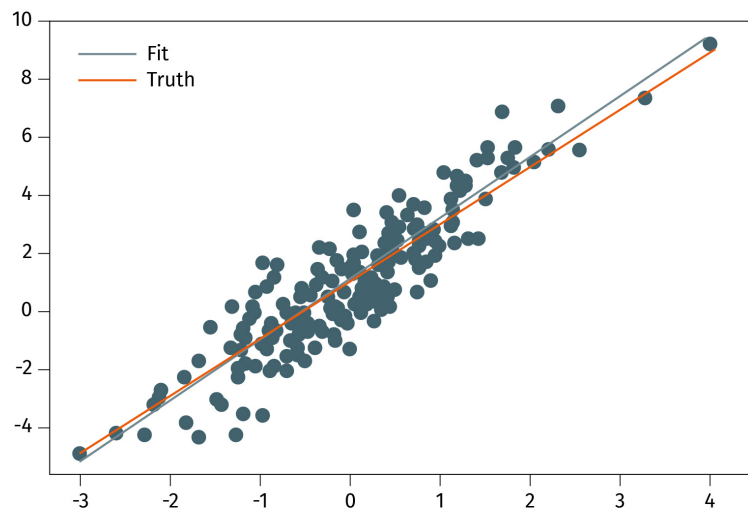
We now use a Bayesian or probabilistic approach to describe the data, again using a simple linear model. We start by saying that our independent variable X with values x is a random variable that has a linear relationship with the independent variable Y with values y. The variables X and Y are connected with some Gaussian noise.

This can be expressed in the language of statistic as

$$Y \sim \mathcal{N}\big(\beta \cdot X + b, \sigma^2\big) \tag{1.30}$$

meaning that we use a linear model with Gaussian noise.

Since we are now focusing on Bayesian probabilistic modeling, we need to assign a prior to all our parameters that need to be determined in our fit. In the case of our simple linear model, we need priors for the slope and intercept of the linear model, as well as the variance of the noise term. If we have any domain-specific knowledge, we can choose the priors accordingly. In this simple example, we choose a wide Gaussian distribution centered around zero as prior for the intercept and the slope. The conjugate prior for the variance would be an inverse gamma distribution. However, it is recommended to use a half-Cauchy distribution as prior for σ (Gelman, 2006; Polson & Scott, 2012).

Using a probabilistic modeling framework such as PyMC, we can use for example the NUTS algorithm to construct a Monte Carlo Markov Chain for this model. First, we check that the has chain converged by determining the $\widehat{R}$ metric. This number should be smaller than 1.01 (Gelman & Rubin, 1992; Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2019).

In contrast to the "standard" linear regression, we now have a full posterior distribution for the model parameters "intercept" and "slope" (beta) as shown below. Note that the shape of the posterior distributions is quite similar to (but not exactly the same as) a Gaussian distribution. The mean values for the intercept (1.0) and slope (2.1) are very close to the true model values we have used to generate the data with and in this case identical to the ones we have obtained using the standard linear regression approach.
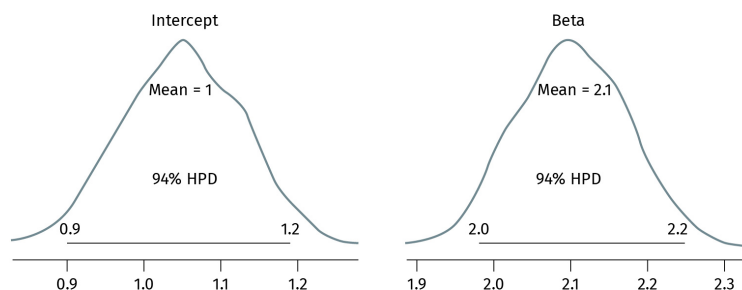
We can also show the Bayesian credible intervals for these parameters.
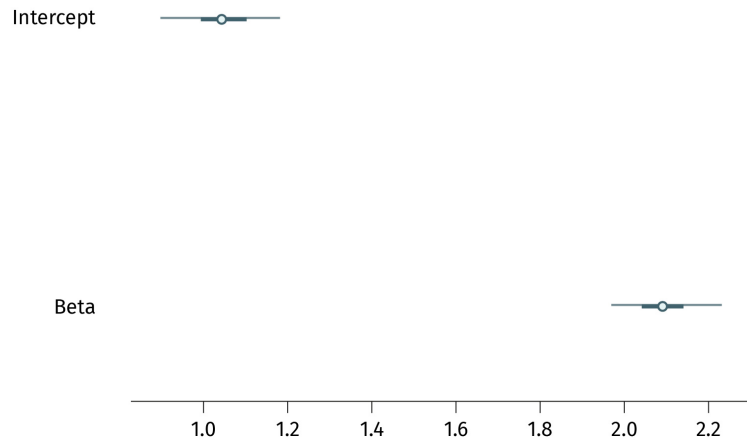
Using the different values from the posterior distributions for the model parameters, we can then draw a representation of all possible models compatible with the data as shown below. Each option is drawn as a light grey line, and the intensity of the black area indicates the most probable model parameters. Additionally, the models for one standard deviation are drawn in orange, and the true model has been used to generate the data with in red.
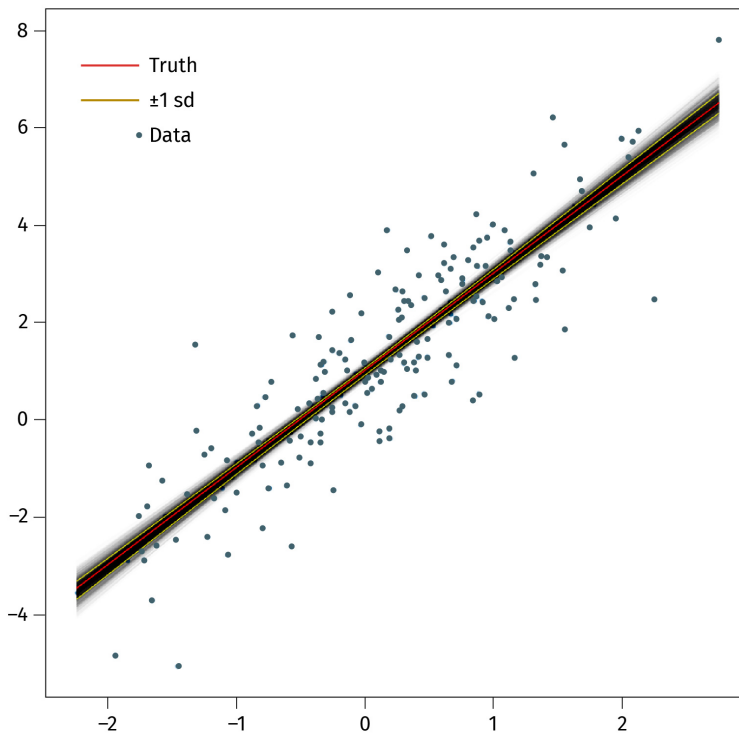
Since we have the full posterior information available, we can not only determine the credible intervals for the model parameters and their posterior distribution but calculate the posterior distribution for each value y of the dependent variable Y at each point x of the independent variable X. For example, at the point $x = 0.2$, the standard linear regression would yield $y = 2.1 \cdot 0.2 + 1.046 = 1.466$ (indicated by the red dashed line), whereas we obtain the posterior distribution for y at $x = 0.2$. The posterior distribution is again quite similar to a Gaussian distribution, and its mean agrees with the result from the ordinary least squares approach.
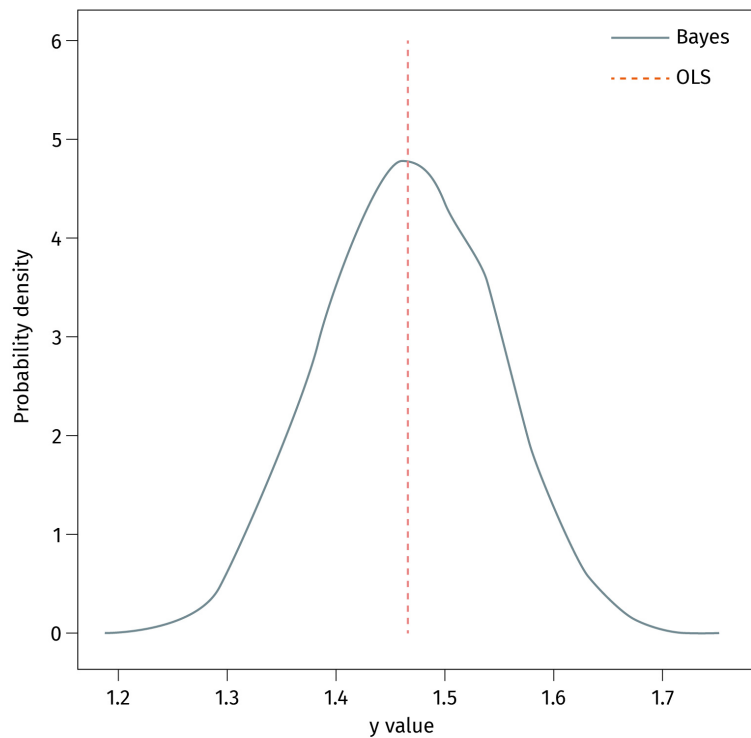
📖 **SUMMARY**

Statistical inference allows us to make statements about systems or their future behavior. One key aspect is that the inferred outcomes are expressed as probabilities. Bayes' theorem is at the core of the Bayesian approach to inference that aims to calculate the posterior probability of a given outcome, taking both the observed data as well as any prior information into account. This prior needs to be carefully chosen and a minimally informative prior should be taken if only very limited information is available. From a practical point of view, in many cases, conjugate priors allow us to calculate the posterior much easier.

In statistical inference, we are often concerned with the analysis of few a quantities. In many situations, however, we need to investigate larger systems described by many variables. Bayesian networks allow to model the dependencies of large systems and derive the predicted outcome based on a large set of input values.

Finally, in probabilistic modeling, we take into account that our model variables themselves are random variables, and we can use Bayesian approaches, in particular a prior, to derive the most likely outcome. The key challenge of probabilistic modeling is that we need to determine the posterior distribution without being privy to all details. This can be done using Markov Chain Monte Carlo methods, and popular algorithms include the Metropolis and Metropolis- Hastings algorithm

HERE YOU CAN FIND THE VIDEOS FOR THIS UNIT.

Source: http://www.kaltura.com/tiny/3bnkr

Alternative text:

UNIT 2

INTRODUCTION TO CAUSALITY

2. INTRODUCTION TO CAUSALITY

Introduction/Case study

Perhaps one of the most important questions regarding the study of causality and causal effects is why one should study it at all. In particular, spectacular progress in the area of machine learning and artificial intelligence has made applications possible that were unthinkable even a few decades ago. Computers and AI systems play the complex game of Go better than humans (Silver et al., 2016) and can detect skin cancer with a level of accuracy on par with human experts (Esteva et al., 2017). However, these systems cannot answer questions as to why something is happening. In many situations, understanding the underlying causal structure may not be necessary. We can improve clinical care significantly if we can quickly and reliably identify skin cancer and use an AI-based system as a diagnostic tool. Similarly, operational procedures (such as replenishing goods at a supermarket) depend on many different factors that influence customer demand. However, we do not have to understand the causal reasoning of individual customers-–rather, we only need to describe the overall effect, i.e., the resulting expected demand of all customers on a given day in a particular store. Additionally, many of these cases where AI based solutions are deployed successfully are within the remit of "narrow AI," i.e., a particular, singular task that the system focuses on. This often implies that causal structures can be explained in terms of domain knowledge, and that many of the correlations the machine learning model relies on are closely related to causal relationships, as they are taken from a very specific application domain.

On the other hand, many questions cannot be properly addressed and answered by this approach. For example, while we may use an AI system as a diagnostic tool, said system cannot answer certain questions, such as whether a given medication really is the cause of the improvement seen during treatment. In particular, there may be many biases present in the data that originate, for example, from common causes between variables.

Correlations between variables found in the data may be spurious and only present in the data because the process of acquiring the data was flawed: had we obtained the data correctly, some variables would have been independent. But as the process of collecting the data was flawed, a later statistical or machine learning model may pick up on these spurious correlations and generate misleading results. Furthermore, some data may be impossible to obtain, for example, due to prohibitive costs or because the means of obtaining the data would be unethical. In many of these cases, careful, causal studies can help to gain deeper insights, allowing us to look at which part of the story the data are not telling.

## 2.1 Correlation versus Causation

We humans are intimately aware of the concepts of both causality and correlation and apply them in our daily lives. Unfortunately, however, our intuition is often wrong. While our behaviour is understandable from an evolutionary perspective, the conclusions and

actions we draw from them are often wrong and we are prone to a wide range of cognitive biases. For example, in his research, psychologist D. Kahnemann has found that our mind can be divided into two cognitive systems through which we experience the world: "system 1" and "system 2". System 1 acts subconsciously and constantly evaluates our world, trying to make sense of it. Only when this is no longer sufficient is system 2 engaged. System 2 is associated with abstract cognitive processes, problem solving, and deliberate thought. Interestingly, system 1 always constructs a causal story based on what we experience (Kahneman, 2012, p. 75). For example, after hearing the following sentence "After spending a day exploring beautiful sights in the crowded streets of New York, Jane discovered that her wallet was missing" (Kahneman, 2012, p. 76), a study found that people associated the word "pickpocket" more strongly with the story than "sights," even though the sentence makes no mention of "pickpocket" or "thief". However, because the sentence is set in New York, we "jump to the conclusion." That is, our system 1 builds a probable and believable causal story that the wallet was stolen rather than lost.

Further studies show that we already have an impression of causality from birth (Kahneman, 2012, p.76), even though most of our everyday causal reasoning happens subconsciously.

When looking at a graph such as figure below, we immediately notice that the two curves follow the same pattern. Even (or especially) when we do not look at the description or details of the graph, we notice that the behaviors are the same and our system 1 intuitively assumes that there is a causal connection between the graphs. Looking at the graph more closely, we see that the two curves show the number of sociology doctorates in the United States compared to the number of non-commercial space launches (Office, 2011; Foundation, 2018). The curves show an obvious relationship––but in thinking more about it (engaging system 2), we cannot possibly find a reason why the number of doctorates awarded in the field of sociology in the United States should be related to worldwide, non-commercial space launches. Admittedly, the graph employs some bad data visualization techniques (such as two scales for the y-axes), and the range of y values is also tuned to make the graph more convincing visually. However, the behavior is real and observable—we say that the two quantities "number of US sociology doctorates" and "worldwide, non-commercial space launches" are correlated.

Figure 7: US Sociology Doctorates versus Worldwide, Non-Commercial Space Launches
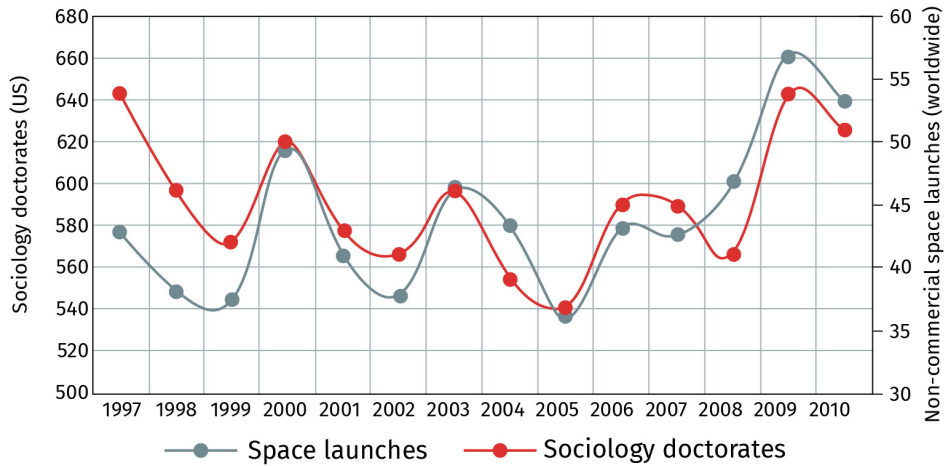
The Pearson correlation coefficient is defined as

$$\rho_{x,y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} \qquad (2.1)$$

where σ_x is the standard deviation of variable x (and correspondingly for y) and $\mathrm{cov}(x,y)$ is the **covariance** of the variables x and y defined by

Covariance
The covariance measures how the variables "covary," i.e., how one variable changes when the other changes.

$$\mathrm{cov}(x,y) = E\big[(x - \mu_x)(y - \mu_y)\big] \qquad (2.2)$$

where $E[\,\cdot\,]$ denotes the expectation value and μ_x is the mean of variable x (and correspondingly for y). The correlation coefficient is normalized using the standard deviations and has a range between $-1 \leq \rho_{x,y} \leq 1$. A value of $\big|\rho_{x,y}\big| = 1$ means that the variables are related 100%, $\rho_{x,y} = 0$ means that they are unrelated. Positive values indicate that the two variables change in the same direction, e.g., if x increases, y increases as well. Negative values indicate that x and y change in opposite directions. Note that the Pearson correlation coefficient only captures linear correlation between variables, as illustrated by figure below.

Figure 8: Correlation Coefficient
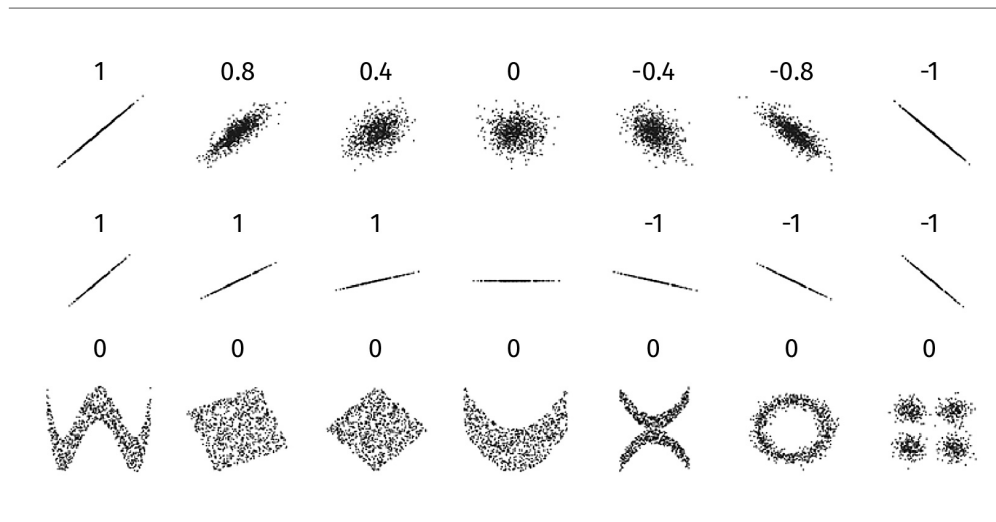
Coming back to the example of sociology doctorates and space launches, we can say that the variables are highly correlated—but we cannot conceive of a reason why there should be a causal relationship between them: correlation does not imply causation. What do we mean by this? Implicitly, we assume that because two variables are correlated and co-vary in a defined way, there is also an underlying cause for this. This is, indeed, one of the main assumptions of statistical modelling, machine learning models or artificial intelligence applications: the model determines the best combination of the input variables or features to derive a prediction for the future behavior of the variable or target of interest. Spectacular successes of such systems (such as the detection of skin cancer with human-level precision (Esteva et al., 2017)) prove that relying on correlations to predict the outcome of a variable works very well. This opens the question as to whether the study of causation is merely a "luxury" or irrelevant in practice. The answer is, of course, no. Studying the causal structure allows us to address questions we cannot answer by looking at the data alone. In other words, the data only contain part of the story.

Why, then, do machine learning or AI models work so well if they do not include causal relationships? The answer lies in the data we give these systems. Using our expert knowledge, we feed a curated set of data and features to the machine learning model or AI system. From these data, it learns about the relationship we are interested in. However, we already know that such a relationship can be extracted from the data and is relevant to the problem at hand. However, if we just pass all data to a self-learning system, we will quickly discover that the resulting predictions will be sub-optimal (or even wrong). Imagine we were modeling the number of space launches and, following the often repeated mantra "it's all in the data" or "the data speak for themselves", we would include the number of doctorates in sociology as a further feature or variable. The subsequent modelling stage will pick up on the correlation and, given that this relationship holds over a long period, might even improve the model. Since there is no causal relationship between the number of doctorates and the number of space launches, our prediction model will lead to incorrect results as we have trained it to include the spurious correlation should the observed correlation no longer hold.

Correlations between variables can also be very misleading. Pearl et al.

(Pearl, Glymour, & Jewell, 2016, p. 3) discuss the example of the variables "cholesterol" and "exercise" From our general knowledge, we know that exercise is beneficial to our health and it is better for cholesterol levels to be low. However, when we plot a (fictitious) dataset containing these two variables, we find that cholesterol and exercise are strongly positively correlated (see figure below, part a).

Figure 9: Correlation Depending on "Hidden" Variables
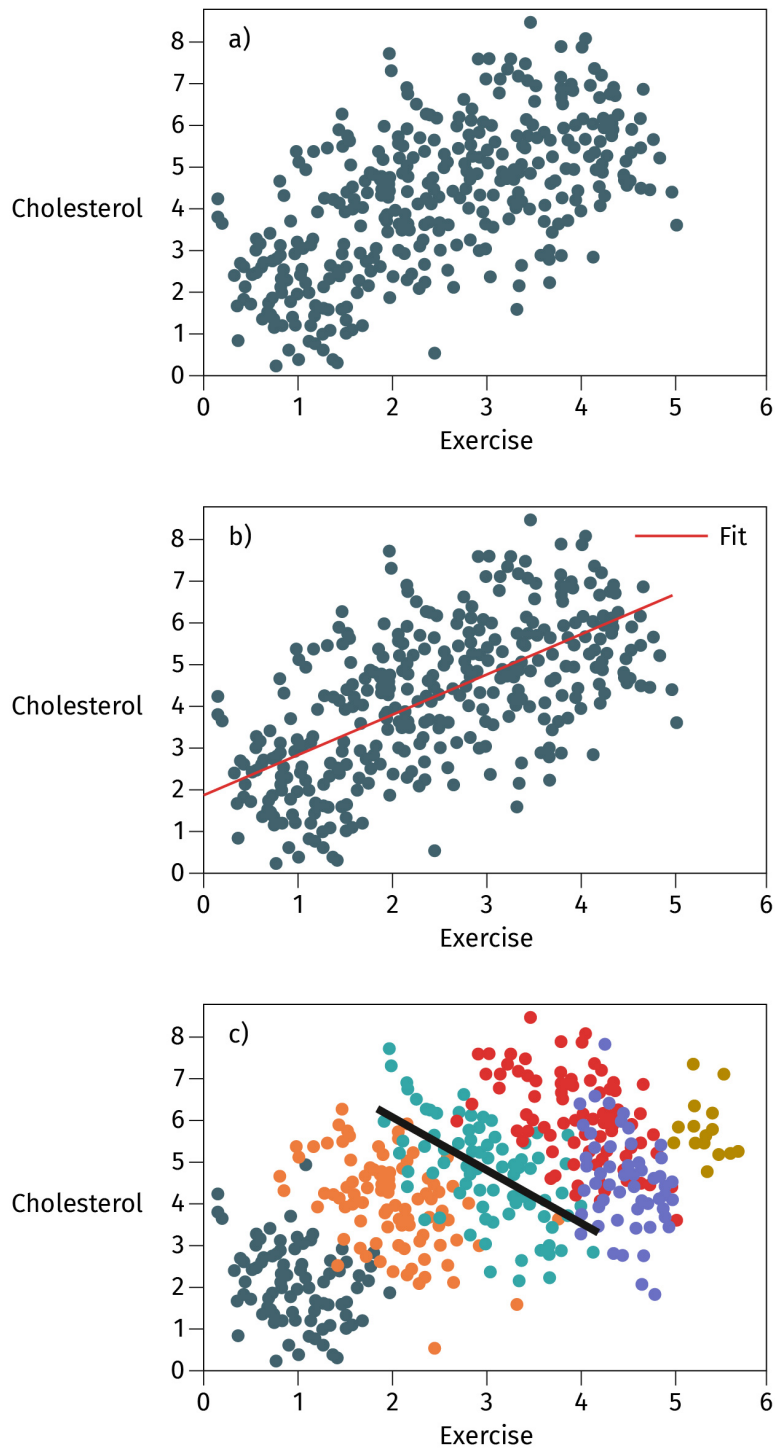
Performing a linear regression (part b) with the functional form $y = mx + b$ results in a slope of approximately 1. From this, we (or a machine learning or AI model) would learn, that an active lifestyle with lots of exercise is associated with high cholesterol levels. In other words, exercise would be bad for our health as it increases cholesterol levels. However, we also know that exercise is good for our health. We cannot answer this conundrum from these data alone. To understand the data, we need to look deeper, and, in this case, it turns out that age plays an important role. If we segregate the data by age (part c), we find the expected (negative) relationship between cholesterol level and the amount of exercise. The "age" variable is a common cause for both cholesterol and exercise: older people tend to have higher cholesterol levels regardless of their level of exercise. If we had given the "age" variable to a sophisticated machine learning algorithm in conjunction with "cholesterol" and "exercise," it might have learned this relationship. However, we would have had to understand the story behind the data in order to know to include this variable in the first place. Also, simply adding all variables at hand into such an algorithm just increases the chance that such misleading correlations are learned, as the machine learning algorithm cannot learn the causal relationship from the data. It "just" exploits the observed correlations optimally. However, it should be noted (as we will learn later) that segregating variables by a third variable (in this example, by age) does not always result in the correct answer. We need to understand the causal story behind the data to decide whether segregating will give us a useful answer or make things worse.

But what if (for some reason) the correlation between, for example, the number of doctorates and space launches was not spurious? What if it actually held? Testing and evaluating such a question is the main focus of causal models. The data cannot answer these questions. We can observe a correlation, but, without further knowledge, we cannot decide whether it is spurious nor real. Admittedly, we would likely not investigate the relationships between these two variables in practice. However, the same underlying question is very relevant to many cases. As an example, in the 1950s, many studies focused on the relationship between smoking cigarettes and lung cancer (Mendes, 2014). Prior to the 1900s, lung cancer was very rare, even accounting for the difficulty in diagnosis then. As smoking cigarettes became more popular and widespread, the number of lung cancer cases started to rise sharply, from 54 cases in 1900, to 4,345 by 1963. Lung cancer became one of the most common types of smoking and lung cancer rose at the same time, it is not the same as proving that smoking is indeed the cause of the increase in lung cancer. Tobacco companies have a strong interest in not establishing a causal relationship, as this would likely result in public policies limiting or banning smoking (as it has been done in many countries much later).

From the 1950s onwards, many studies were performed that proved a causal relationship between smoking and lung cancer. However, this topic also highlights some issues with experimental studies: many studies are performed as **randomized controlled trials** (RCT). Since each individual in such a study is assigned to the treatment or control group randomly, there is no cause that could influence whether that individual would receive the medicine. In a medical study, we can then determine if the medicine works by observing the outcome in the two studies. However, what if administering a medicine (or withholding it) were unethical? What if we forced participants to take a substance we suspect is lethal? In the case of the tobacco studies, the researchers used volunteers. This might solve the ethical dilemma, as the participants do smoke voluntarily. Even though they

Randomized controlled trials
In a RCT, subjects are randomly assigned to the treatment or control group.

were made aware of the risk, however, relying on volunteers might introduce a bias, since only a specific type of person might volunteer. Again, we need to understand more of the causal story behind the data to understand possible complications.

Thus far, we have pointed out that causal relationships play an important role when understanding the relationships between variables and their values.

> 📖 **CAUSAL RELATIONSHIP**
>
> "A variable X is a cause of a variable Y if Y in any way relies on X for its value." (Pearl et al., 2016, p. 5)

Pearl et al. use the example of listening to illustrate the definition: "X is a cause of Y if Y listens to X and decides its value in response to what it hears." (Pearl et al., 2016, p. 5)

2.2 Granger Causality

Granger causality (Granger, 1969) is a concept of causality that is exclusively related to the analysis of **time series** data, and hence quite different to the discussion of causality in the remainder of this text. See, for example, Eichler (2012) for more information on this topic. It was developed in the context of economics and therefore caution should be exercised when Granger (or G-) causality is used for other time series data. In this context, causality is defined temporally: a preceding event can cause a later one, but a later event cannot cause an earlier one.

Time series
Time series data are data that follow a distinct temporal order.

If we have two (or more) time series, we can intuitively understand Gcausality in the following way: a specific feature of one time series causes feature in the other time series at a later time, i.e., at a given lag. This is illustrated in figure below (BiObserver, 2014), where the big spikes in the time series X occur in time series Y at a later time.

Figure 10: Granger Causality (BiObserver (Wikipedia) CC BY-SA 3.0)
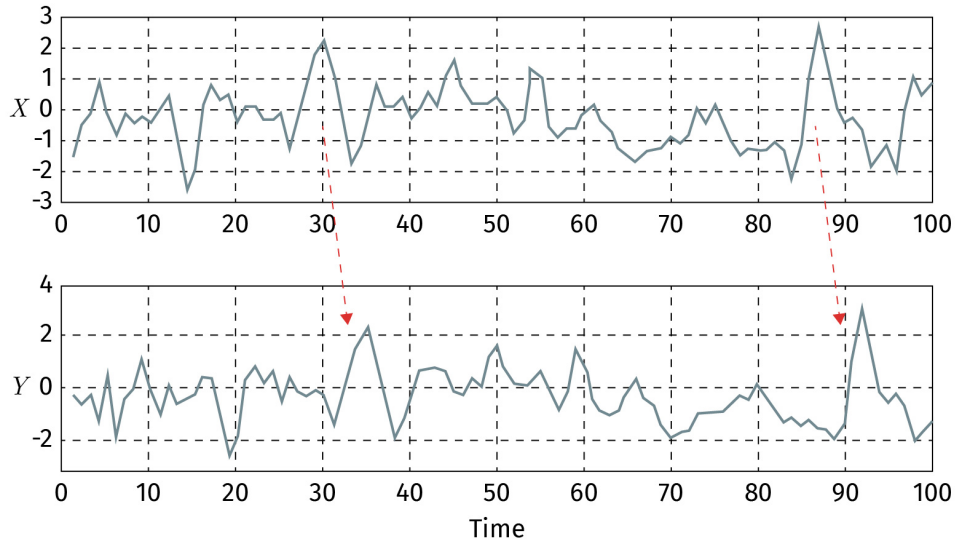
This means that one time series (X) contains some information that can be used to explain the behavior of the other time-series (Y). Hence, including X in the forecast of Y leads to better predictions than when X is not used.

A bit more formally, we can look at two time series X_1 and X_2, where we use the subscripts to indicate that we could extend the argument to more time series up to some index X_n. We can then write the system of time series equations for an auto-regressive model for the two time series as:

$$X_1(t) = \sum_{i=1}^{p} \alpha_{11,i} X_1(t-i) \quad + \sum_{i=1}^{q} \alpha_{12,i} X_2(t-i) + \epsilon_1(t) \tag{2.3}$$

$$X_2(t) = \sum_{i=1}^{p} \alpha_{21,i} X_1(t-i) \quad + \sum_{i=1}^{q'} \alpha_{22,i} X_2(t-i) + \epsilon_2(t) \tag{2.4}$$

Lag
The lag determines the time shift between the original and modified time series.

where coefficients α determine the strength with which each **lag** i contributes to the time series, and the order of the auto-regressive model is given by p, p', q, q'. The numbers $\epsilon_1(t), \epsilon_2(t)$ are residual uncertainties. As we can see, the two time series depend on each other. For example, if $\alpha_{12,i} \neq 0$, then $X_1(t)$ depends on $X_2(t)$, and vice versa for $\alpha_{21,i} \neq 0$. We say that the time series are connected via G(ranger)-causality if we can establish that $\alpha_{12,i} \neq 0$ with some level of significance.

Note that this implies that both time series X_1 and X_2 can be connected by Granger causality: X_1 can be the cause of some feature in X_2, and, at the same time, X_2 can be the cause of some other feature in X_1. As mentioned before, we can extend this to n different time series $X_1, ..., X_n$ that can all, to some varying degree, be connected by Granger causality with each other. The dependencies can also be expressed as directed acyclic graphics (Chen & Hsiao, 2010) which allows to connect Granger causality to the concepts developed in the rest of this textbook.

In the considerations above, the lag at which a feature in one time series causes the feature in the other time series was always fixed. This means that when the feature in the time series occurs that causes the effect in the other time series, the effect would appear there with a fixed delay. In many systems, this delay is not fixed, but may be variable, which can be included in an extended definition of Granger causality (Amornbunchornvej, Zheleva, & Berger-Wolf, 2019) When looking at the temporal order of events, it is important to avoid the "post hoc ergo propter hoc" (Latin for "after this, and, therefore, because of this") fallacy. Events can occur after one another, even though they are not causally related. For example, the rooster crows at sunrise but does not cause the sun to rise. Examples of this from the medical field are given in (Grouse, 2016).

2.3 Directed Acyclic Graphs (DAG)

Causal relationships can be expressed in a number of ways. A very powerful method is centered on a graphical representation called "directed acyclic graphs" (or DAGs), and we will follow the notation in (Pearl et al., 2016).

Figure 11: Basic Graphs

a) Undirected graph

b) Directed graph
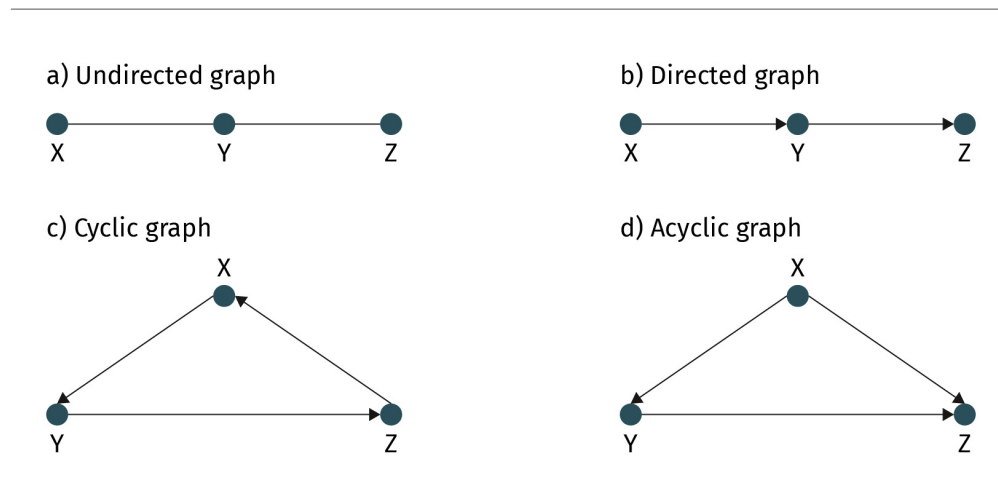
c) Cyclic graph

d) Acyclic graph

Figure above shows the basic setup of simple **graphs**. In this example, the nodes are labelled X, Y, Z. The edges connecting the nodes are either undirected (see part a) or directed (part b). In the latter case, we use little arrows to indicate the direction. Hence, directed edges come out of one node and go into another. The node from which the arrow emerges

Graphs
A graph is made from "nodes" or "vertices" that may or may not be connected to other nodes via "edges."

is called the "parent," and the node into which the arrow enters is called the "child." The relationship between grandparents and grandchildren is defined accordingly. Occasionally, we may use undirected edges as shown in part a) of figure above to indicate that two variables are associated with each other and are therefore correlated, but we do not know which variable is the parent and which is the child. However, the point of causal diagrams is we will typically draw causal graphs for all alternatives and then find ways to experimentally test which is the correct graph.

A path between any two nodes is called a directed path, if we can follow the arrows emerging through the parent, going through all children and grandchildren until we arrive at the destination node.

If any two nodes are on a directed path, the first node is called the ancestor, and all subsequent nodes on the directed path are the descendants of this node.

In a cyclic graph (part c), we can return to the origin node following a directed path. This means that, starting at one node, we can follow the direction of the arrows and come back to the node we started from. In an acyclic graph (part d), no such directed path exists, and we can move from ancestors to descendants—but not back.

It is important to note (and remember) that information can "flow" along any edge (directed or undirected) and, in case of a directed edge, even against the direction of the arrow. Although it may appear counterintuitive or confusing at the beginning, an arrow does not indicate that information only flows in the direction of the arrow.

Causal graphs are mainly constructed from, and are extensions of, part b and d of figure above. In these graphs, the arrow indicates a causal relationship: the ancestor can causally influence the descendants (in the direction of the arrow, but not the other way round). As an example, we can imagine a barometer. If it rains, the needle will point to a low value: rain $\rightarrow$ barometer value. The underlying physical explanation is that the barometer measures the air pressure, and during bad weather, the air pressure is low. For the point of illustrating the causal relationship, we can imagine that the variable "barometer value" listens to "rain" (as proxy for atmospheric pressure). However, even though we can manually force the needle of the barometer to any value, the weather will not change. Causal graphs are typically read either left to right or top to bottom. This is not a strict rule, but we generally try to arrange the graph such that the ancestors are either toward the left (or top) of the graph and the descendants towards the right (or bottom).

part d of figure above also illustrates that we have generally two types of effects:

a direct effect and an indirect effect. The variable X causally affects the variable Z directly —this is called the direct effect. In addition, X also affects Y, and Y in turn affects Z. Hence, even in the case where there is no direct effect from X to Z, X can still affect Z via Y—this is called the indirect effect. An important path of building causal models is to add all the ways the variables can influence each other and in which way they are (causally) connected.

Variables can also have more than one cause. For example, in part d of figure above, the variable X is a common cause of both Y and Z and (causally) influences both. X, Y, and Z are associated with each other.

By this we mean that these variables are related to each other but either we do not know their relationship yet, or we do not make such a statement.

More formally, we can say that two variables are associated when observing one changes the probability of observing the other. This implies that the variables are correlated but we do not want to make a causal statement: The association can be due to a causal connection between the variables which also makes them correlated. This association can also originate from, for example, a spurious correlation because we have not (yet) taken all causal dependencies into account. By saying that a set of variables are associated we want to express that we know that they are related to each other in the sense that observing one changes the probability of observing the other(s), but we do not want to make a further statement as to why the variables behave this way.

In the above example, neither Z nor Y can be a cause for the behaviour of X. Expressing this using the earlier definition of causality, the variables Y and Z listen to the value of X to define their values—but X does not listen to either Y or Z. Nodes that are not connected via an edge are not associated with each other, hence we neither have a causal relationship nor can we observe a correlation between them—they are independent.

As an example, consider figure below, part a.

Figure 12: Common Cause
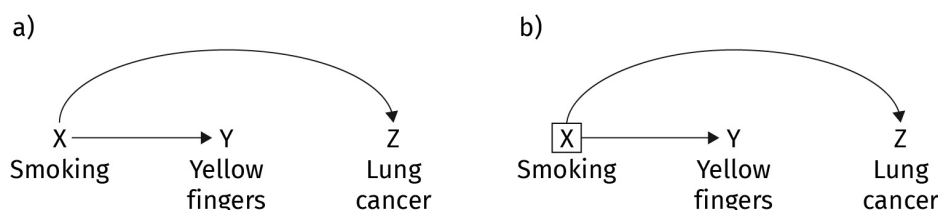
Smoking (variable X) is both a cause to yellow fingers (Y) and lung cancer (Z). Note that there is no arrow from Y (yellow fingers) to Z (lung cancer). Having yellow fingers does not cause lung cancer. Note that we did not use any data to construct the causal graph. Instead, we constructed the causal graph from the expert knowledge we may have in this area. However, if we were to look into a (large) dataset, we would observe the following: the variables Y (yellow fingers) and Z (cancer) are associated, i.e., the proportion of individuals that are affected by lung cancer is different for those with and without yellow fingers. Hence, we observe an expected correlation between yellow fingers and cases of lung cancer. However, this correlation is not due to the fact that yellow fingers are a cause of lung cancer, but that both yellow fingers and lung cancer share a common cause: smoking. Hence, two variables can be associated even though there is no direct connection (an edge) between them in the causal graph. This constellation can lead to a bias in the analy-

sis and illustrates that the information flows both in the direction of the arrows and against it: Informally, we can say that yellow fingers and lung cancer are associated because the information about lung cancer flows backwards via the common cause into the variable describing yellow fingers. We call the variable X (smoking) a confounder of variable Y and Z.

So far, we have only considered all values of the variables. We now look at the relationship between Y (yellow fingers) and Z (lung cancer) in different slices of the common cause X (smoking). In this case, X is a **binary variable**, and we consider both options, smokers and non-smokers, and look at the association between Y and Z for each value of X. This is called conditioning, and we indicate that we condition on a variable by adding a little box around it in the causal graph, as shown in part b of figure above. We then look at the data and check the association between yellow fingers and cases of lung cancer, for example across all individuals who never smoked. Since yellow fingers were associated with smoking and we are now looking at people who have never smoked, learning that an individual has yellow fingers does not change their chances of developing lung cancer. In the same way, if we look at all individuals who smoke and the rate of lung cancer, we find that this rate doesn't change regardless of whether or not the individual has yellow fingers. Hence, in each stratum of the common cause X, the association between the variables Y and Z is removed—even though it is present if we don't condition on X, but look at the complete or marginal distribution. This also holds if X is not a binary variable and can take any range of values.

Binary variable
A binary variable can only take two values, such as "true" or "false."

> ### CONFOUNDER
> We expect an association between two variables, even if the variables are not causally connected when sharing a common cause. This confounding effect is blocked if we control for the common cause (the confounder).

Similarly, causal graphs can contain structures with common effects. As an example, consider the causal graph in figure below, part a.

PREVIEW-PDF, erzeugt: 2024-06-18T14:30:01.73+02:00

Figure 13: Common Effects
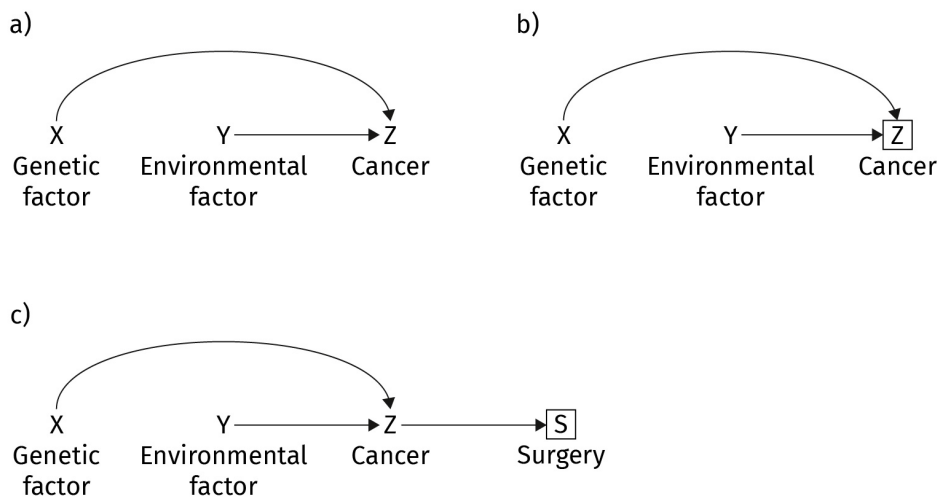
With a simple picture, we can imagine that developing cancer is due to a genetic factor and an environmental factor. For simplicity, we assume that these factors are binary, i.e., either you have a genetic disposition to develop cancer (or not); or you are either exposed to some environmental factor (or not).

There are arrows from both the genetic and the environmental factor to cancer, but there is no arrow from the genetic factor to the environmental factor. Your genetic code cannot causally influence the environment, for example, air pollution. Again, we have drawn this causal diagram using expert knowledge and did not rely on data. If we were to look at a large data set, we would find that X and Y are indeed independent. If two variables have a common effect, they are still independent from each other if they are not causally connected. For example, a given fraction of the population has a genetic factor that raises the chances of developing cancer.

This factor does not vary depending on where the people live. Likewise, being exposed to some environmental factor does not alter your genetic predisposition.

However, this situation changes if we now look at specific values of Z, i.e., individuals who have developed cancer. As before, we indicate that we look at specific values or condition on the variable by adding a square box around the variable, as shown in part b of figure above. Let's say we look at all individuals with cancer, i.e., $Z = 1$. We then look at the values of X and Y. We now find that X and Y are indeed associated, whereas they were not when we looked at all values of Z. We can understand it this way: If a person has cancer ($Z = 1$) and does not have a genetic prevalence ($X = 0$), it is more likely that this individual was exposed to the environmental factor ($Y = 1$). This is not due to any causal connection between the two factors. Rather, since the the person has cancer, it must have been caused by something - and it wasn't the genetic factor. This makes the environmental factor more likely, and X and Y therefore become correlated. This is called the selection bias,

a systematic bias that arises due to the selection of our data set. Note that this applies not only to the common effect, but also happens if we condition on any descendants of this variable, as shown in part c of figure above. In this case, surgery is a treatment for cancer. Since we do not perform the surgery randomly, S can act as a proxy for Z and hence causes the same bias as if we had conditioned on Z directly.

> ### 📖 SELECTION BIAS
> We expect an association between variables that are not associated or otherwise causally connected if we condition on a common effect.

The above discussion highlights that causal graphs representing causal models are closely related to statistical models. For example, if we draw an edge between two nodes, we expect a correlation and a causal relationship.

If no such edge is present, we understand that the nodes are independent from each other.

In some causal models, it is important to account for any exogenous causes that influence the variables. We denote these with U (for unknown), indicating that the observable and measurable variables (such as X, Y or Z) are influenced by external effects that we cannot access or measure directly.

A simple example is shown in figure below. In more complicated settings, unobserved causes may influence more than one variable or lie on a directed path.

Figure 14: Graph with Unobserved Causes
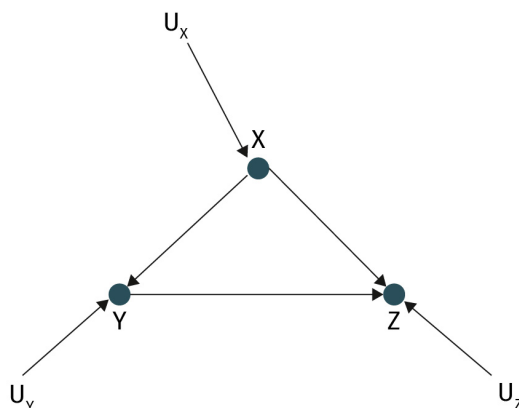
Furthermore, we have assumed so far that all variables measured so far can be measured fully and correctly, i.e., there is no measurement error.

In many real situations, this will not be possible, and we need to account for the fact that the variable we observe or we can measure is not the one causally connected to other variables. This is illustrated in figure below: we are interested in variable Z, or, more precisely, the effect of X and Y on Z. However, in this case, our measurement of Z is impaired by some measurement error, and we therefore only have access to the variable Z^*, which is influenced both by the "true" behaviour of the underlying variable Z and the uncertainty of the measurement process U_Z.

Figure 15: Variables with Measurement Errors

When constructing DAGs to represent causal models, it is paramount to remember that these graphs are built to express testable models—not represent the most realistic way any number of variables might be influenced by any number of effects. We can always find (un-) observable effects that are influenced by further (un-) observable effects, etc. The point is to build a simple model of the causal relationship we wish to explore in our research question. As such, the DAG is a simplification. That being said, it must not be too simple. In particular, we need to be careful to include all common causes of all nodes we add to the graph, as these can lead to biases.

2.4   Elements of Causal Graphs

In the following section, we want to investigate the basic building blocks of causal graphs in more detail. In particular, we will look at how three variables or nodes (say A, B and C) can be connected by arrows or directed edges in directed acyclic graphs (DAGs). In particular, we look at mediators (chains of variables), forks (confounders), and colliders. Each type of connection describes a different way that the variables can affect each other.

2.4.1. Fork or Confounder

The **fork** is a frequently encountered constellation in a causal graph. This is illustrated in figure below, which shows the same constellation in three different ways. In each case, the arrows exit variable B and enter variables A and C.

Fork
In a fork, arrows split into two different directions.

Figure 16: Fork or Confounder
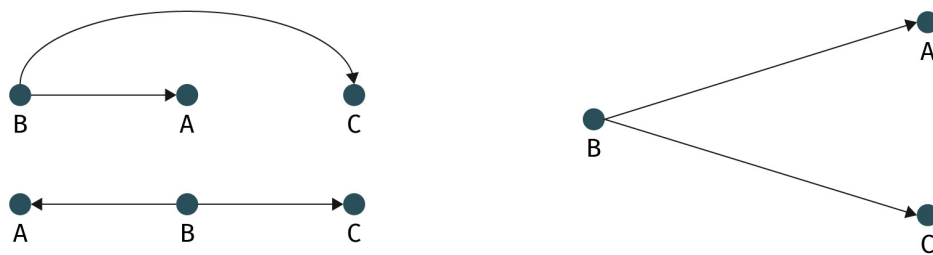
Variable B in the fork is often called a "confounder." We have already previously encountered this situation when we analyzed the relationship between smoking, yellow fingers and cancer. In that case, smoking (B) was a common cause between yellow fingers (A), and cancer (C), and we found that A and C were associated even though there was no causal connection between them. This can be illustrated by a further example (Pearl & Mackenzie, 2018, p. 114): we want to investigate the reading ability of school children. If we look into the data, we find an association between shoe size and reading ability. While the correlation found in the data is observable, we cannot imagine why shoe size should be related to the ability to read. However, there is a common cause to both shoe size and the reading ability: age. On average, older children will have larger shoes and can read better. The corresponding causal graph is: $\text{Shoe size} \longleftarrow \text{Age} \longrightarrow \text{reading ability}$. Here, age is a common cause to both shoe size and reading ability. This is why variable B in the fork (age in this example) is called a confounder: it introduces a spurious association between variables that are otherwise unrelated. Controlling for B, i.e., looking at the values of A and C in separate regions of B, removes this association: If we look at, e.g., only children of a specific age, the ability to read is no longer associated with the shoe size.

For further information also see Pearl and Mackenzie (2018, p. 114) Pearl et al. (2016, p. 35ff).
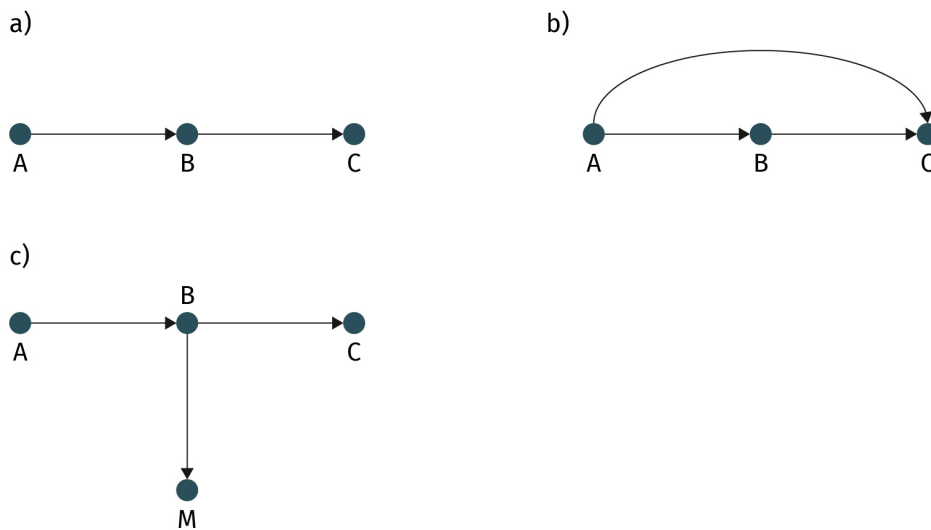
2.4.2. Chain or Mediator

The chain connects the three variables (A, B and C) with directed edges via arrows: $A \rightarrow B \rightarrow C$, i.e., A is the parent of B, B is the parent of C, and A is the grandparent and ancestor of C, as shown in part a of figure below. The variable B in the middle is also called the "**mediator**."

Mediator
The mediator transmits the effect from cause to outcome.

Figure 17: Mediator

We can understand this construct using the example of a fire alarm (Pearl & Mackenzie, 2018, p. 113). Although we call them fire alarms, most detectors work by detecting the presence of smoke and not, for example, heat. The corresponding DAG is Fire $\rightarrow$ Smoke $\rightarrow$ Alarm. Hence, the For example, we could have a fire that produces no smoke (or so little that the detector would not recognize it). In this case, the alarm would not be triggered even if there was a fire. This means that mediator B (smoke) "screens off" the value of the original cause A (fire): once we know the value of B, knowing the value of A will not add any additional information.

Generally, we do not need to build causal graphs this way. For example, if we are interested in establishing whether smoking causes lung cancer, we could just use A for smoking and C for lung cancer. We do not need to know the exact mechanism causing this to happen. Or, if we want to know if some medicine (A) is the cause for the improvement in the outcome (C) of the treatment, we do not need to know the underlying mechanism to establish that the medicine works. Note that this is not the same as understanding why the medicine works—here, we just want to establish the fact that the medicine is indeed the cause of the improvement seen in the patient receiving the treatment (as opposed to some other factor).

However, there are some cases in which we want to use a chain or put a mediator in explicitly. For example, imagine that we know that a medicine works and at least part of it works in a specific way. Then, we can draw the DAG shown in part b of figure above: We know medicine A has an effect on outcome C of the treatment. We then use mediator (B) to separate the effect the medicine has via a specific mechanism from the general effect.

An important question is whether or not we should control for mediators, i.e., look at specific values and add a little box around the mediator in the causal graph to illustrate that we fix its value. In most settings, we do not want to control for the mediator in the chain as

fixing B would "screen off" the value of A, and we would not be able to learn about the causal relationship. For example, if we controlled for smoke in fire alarms and only look at cases $B = 0$ (no smoke), the alarm would never go off, regardless of whether we would also observe a fire. Hence, when we have made the mistake of controlling for the mediator, we might conclude that fire has no causal effect on fire alarms. In practical situations, the mediator is often replaced by a **proxy**, as illustrated in part c of figure above. For example, we might take the affiliation to a religious group as a proxy for religious belief or the membership to a political party as a proxy for political views. In each case, we cannot directly measure the causal variable (B), but we can acquire data that is closely related to it. Typically, proxies are not perfect representations of the variable itself. However, controlling for the proxies can have the same effect as controlling for the mediator itself.

Proxy
A proxy is a variable that is used to measure a variable that is not directly accessible.

However, there are some scenarios where we want to control for mediators.

Suppose we want to establish the causal effect of variable A on C. However, as a complication, both have a common but unknown or immeasurable cause (U), as shown in part a of figure below.

Figure 18: Controlling for Mediators
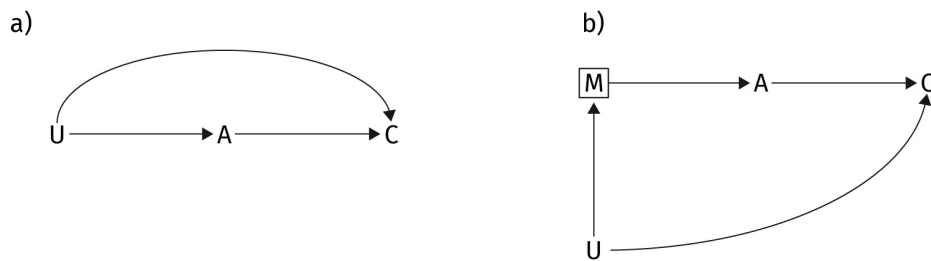
Hence, U is a confounder for the effect of A on C. Unfortunately, since U is immeasurable, we cannot control for it by fixing its value. However, if we can measure a variable (M) that mediates the effect of U on A, we can control for the mediator M instead of the confounder U, thereby determining the causal relationships.

The same holds if M is not a mediator for the effect of U on A, but for U on C.

It is important to note that mediators and confounders share the same independence condition: the causal graphs are given by $A \longleftarrow B \longrightarrow C$ for the fork or confounder and by $A \longrightarrow B \longrightarrow C$ for the mediator or chain. In both cases, conditioning on B, i.e., fixing the value of B makes the variables A and C independent. Hence, we cannot, for example, use data alone to determine which causal structure is correct. Instead, we need to use our expert knowledge and may need to expand the causal graph to determine which is correct.

For further details, see (Pearl & Mackenzie, 2018, p. 113) (Pearl et al., 2016, p. 35ff).

2.4.3. Collider

Figure below shows three examples of how a **collider** could be used in a causal graph.
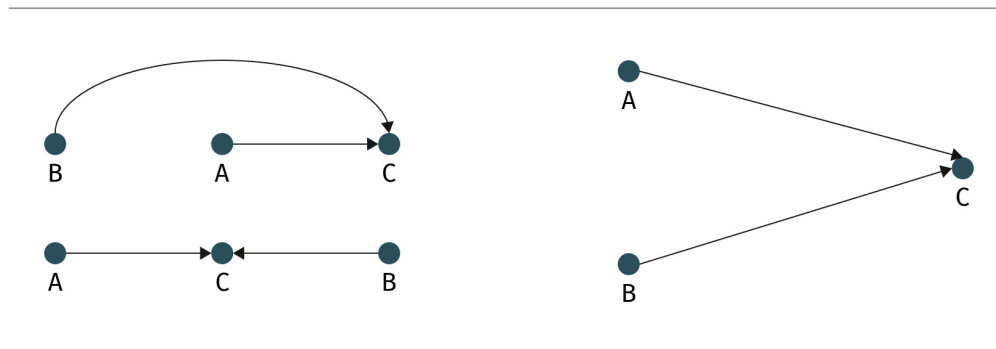
Figure 19: Collider

This construct is used when a variable has multiple causes. In our examples, arrows from both A and B enter C, meaning that A and B are causes of C. We had already encountered colliders when we discussed common effects earlier in the example of cancer that can develop from either a genetic or an environmental factor. This example also highlighted one of the most important aspects of working with colliders: when conditioning on a collider, two variables that are otherwise unrelated become associated.

A further example illustrates this using three aspects of Hollywood actors: Talent $\longrightarrow$ Celebrity $\longleftarrow$ Beauty (Elwert & Winship, 2014). Both talent and beauty contribute to the success of actors. However, in the general population, these factors are unrelated, which means that if we look at a sample of talented individuals, the distribution of the variable "beauty" is not different from the sample where we set "talent = 0." However, if we condition on the collider by setting "celebrity = 1," we find that talent and beauty become associated, even though there is no causal connection between them. This is because we know that a given actor is a celebrity, and if this is not due to talent, beauty must play a stronger role. Therefore, talent and beauty become negatively correlated.

Another way to see how colliders work is to use three variables (X, Y, Z) that are connected via the simple equation $Z = X + Y$; the variables X and Y are independent from each other.(Pearl et al., 2016, p. 41). If we know the value of X, say $x = 3$, we would not know anything about Y, since X and Y are unrelated. However, if we also knew the value of Z, e.g., $z = 10$, then, knowing X, we are able to infer Y. Hence, X and Y become associated if we know Z.

Note that if we condition on a descendant of a collider, this has the same effect as conditioning on the collider itself, i.e., the two variables with arrows pointing into the collider and further on into the descendant become associated when being conditioned on.

For further details, see Pearl and Mackenzie (2018, p. 115) and Pearl et al. (2016, p. 40ff)

2.4.4. Paths

We have so far discussed the basic building blocks from which we construct causal graphs. In particular, the variables are also known as nodes in the graph and may be known or measurable or unknown or not measurable.

These nodes are connected via arrows and elements such as chains (or mediators), forks, or colliders express how the variables are related to each other. Once we build a causal graph, the nodes become connected by **paths**.

Paths
A path is made of a sequence of connections (or edges) between the nodes.

> **PATH**
>
> A path in a causal graph is any route between any two nodes in the graph connected by arrows. Some paths follow the direction of arrows, whereas other paths do not.

It is important to remember that the paths are made from the arrows, but valid paths can go in the direction of the arrows or against the direction of the arrows. This may be a bit confusing at first, as we intuitively assume that a path follows the direction of the arrows—but this is not the case.

Figure 20: Paths in a Collider
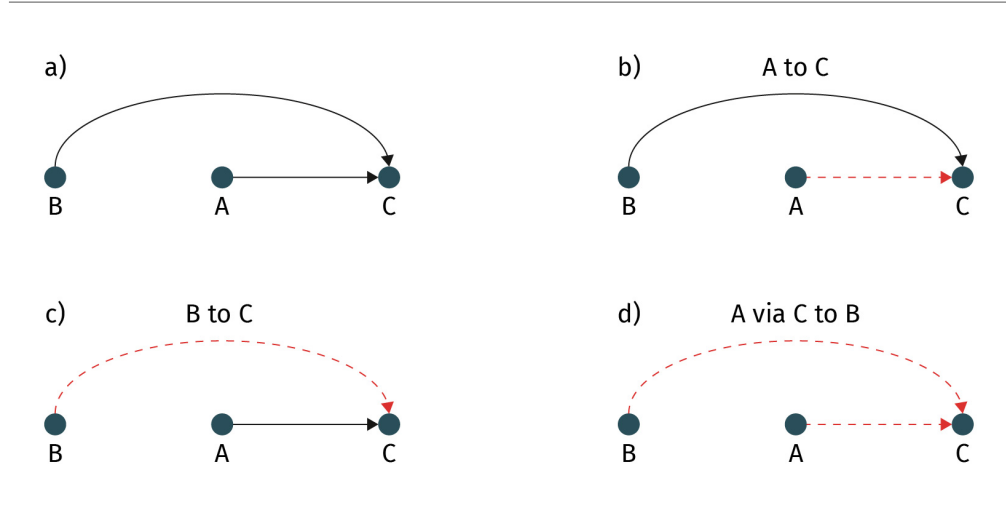
In figure above, we can explore how to determine the paths by using the example of a simple collider where A and B are common causes of C. Part a of the figure shows the collider, and this simple graph has three paths:

1. From A to C in the direction of the arrow (part b)
2. From B to C in the direction of the arrow (part c)
3. From A to C in the direction of the arrow and then from C to B against the direction of the arrow (part d)

A path can be either blocked or open, and to determine whether a path is open or not, we need to look at the behavior of the elements on the path.

We have already seen examples of this when we looked at the association between variables. When we looked at the collider in terms of common causes, we saw that, for example, the variables "talent" and "beauty" become associated when we know that an actor is a celebrity, even though "talent" and "beauty" are unrelated amongst the general population. We can also say that the path between "talent" and "beauty" is blocked by the collider but becomes open when conditioning on the collider (Pearl et al., 2016, p. 46).

> **PATH RULE FOR COLLIDERS**
>
> A collider blocks a path and, hence, the association between variables along the path they lie on. Conditioning on a collider opens the path, and the variables become associated. This also holds for conditioning on descendants of colliders.

The opposite is true for chains and forks. In the case of forks, we have already seen this in the example where smoking is a common cause of both yellow fingers and lung cancer. Even though yellow fingers are not a cause of lung cancer, the two are associated because they share a common cause (smoking). In this situation, the central element of the fork is also called "confounder." We then saw that we can remove this spurious association by conditioning on the confounder in the fork. We can express this in the following rule that holds for all non-colliders (Pearl et al., 2016, p. 46).

> **PATH RULE FOR NON-COLLIDERS**
>
> A path through a non-collider (fork or chain) is open, meaning variables on a path connecting any two nodes are expected to be associated.
>
> Conditioning on the non-collider (e.g., on a confounder or mediator) blocks the path and the association is removed. This also holds for descendants of non-colliders.

Note that colliders are path specific. If multiple paths go through a node, that node may be a collider for some of the paths but not others.

Figure 21: Paths in a fork
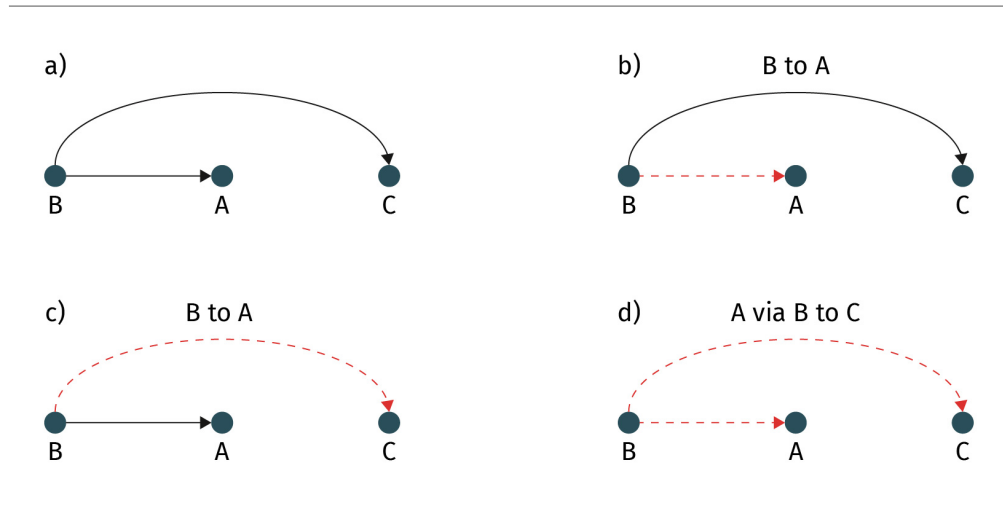
2.5 D-Separation

Causal graphs, or DAGs, are a model of how we think a number of **variables** do or do not depend on each other. If we think that a variable (X) has a causal influence on another variable (Y), we draw an arrow from X to Y. This also means that we expect to observe an association or correlation between these variables in the data. If there is no causal influence of X on Y, we do not draw an arrow. However, as we have already seen, this does not generally mean that that we do not expect to observe an association between X and Y in the data. In several examples, we have already seen that X and Y may be associated due to the structure of the graph. Depending on the arrangement of colliders, chains and forks, as well as whether we condition on some of these elements (or not), we expect to observe or remove an association between variables. Paths allow us to traverse the graph and determine the relationship between any two variables, even if they are far apart in the graph.

As we have discussed before, a causal graph is a tool that helps us understand a concrete research question or a specific relationship. It is not intended as a "world-model" that explains everything that may be connected to some variable we come across in any given question. As we pointed out earlier, we can always find possible causes of causes of causes and so on. Attempting to build such an inclusive diagram quickly leads us down the proverbial rabbit hole. Instead, we focus on a specific question such as "Does smoking cause cancer?" or "Is medicine X a cure for disease Y?" or "Does the vaccine work?" We only need to include the variables that are relevant to the question at hand. In most cases, we do not even include mediators in chains unless we are interested in their specific properties.

However, we do need to make sure that we include all relevant variables, all common causes to each variable, as well as unobservable variables that influence one or more variables in the graph.

One of the most important applications of the resulting graph is determining whether we are prone to any bias in the analysis by measuring certain variables in the data, i.e., which data we take and which variables we condition on. Remember that conditioning on variables can introduce an expected association or remove them, depending on how the variable is connected to other variables in the graph. We have previously discussed that we can analyze the path between any two variables or nodes in or against the direction of the arrows connecting them to determine whether we expect an association.

Unfortunately, variables can be connected by multiple paths in a complex graph, and along each path, there will be any number of colliders, forks, or even chains. To be able to analyze a graph, we need a criterion that determines whether we expect an association between two variables in the data and if there is a way to remove unwanted associations that lead to biases or if handling a specific variable introduces a new bias. This criterion is called "**d - separation**" (where d stands for "directional") (Pearl et al., 2016, p. 46). This means that variables (nodes in the graph) X and Y are d-separated when either there is no path between them (i.e., no arrow along the path) or if all paths between them are blocked. If even one path is not blocked, the variables are d-connected, and we can expect an association in the data. Pearl uses the example of pipes (Pearl et al., 2016, p. 46) that represent the paths. If all pipes are blocked, water cannot be exchanged between the nodes. If at least one pipe is open, water can flow.

D - separation
Any two nodes are called "d - connected" if there is a path connecting them and "d - separated" if there is no such path.

As with pipes, paths only need to be blocked in one place.

As we have discussed before, the following nodes can block a path (Pearl et al., 2016, p. 46):

- a collider that is not conditioned on (or any of its descendants)
- a chain or a fork that is conditioned on

More formally, we can define d-separation as the following.

📖 **DEFINITION OF D-SEPARATION**

Let p be a path between nodes. Then p is blocked by a set of nodes Z if and only if (Pearl et al., 2016, p. 46): 1. The path p contains a chain of nodes $A \longrightarrow B \longrightarrow C$ or a fork $A \longleftarrow B \longrightarrow C$ such that the middle node B is in Z, i.e. that B is conditioned on; 2. The path p contains a collider $A \longrightarrow B \longleftarrow C$ such that the collision node B (or any of its descendants) is not in Z, i.e., that neither B, nor any of its descendants, is conditioned on.

Set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y (Pearl, 2009, p. 17).

As an example, consider the causal graph in figure below (Pearl et al., 2016, p. 47).

Figure 22: A Complex Causal Graph
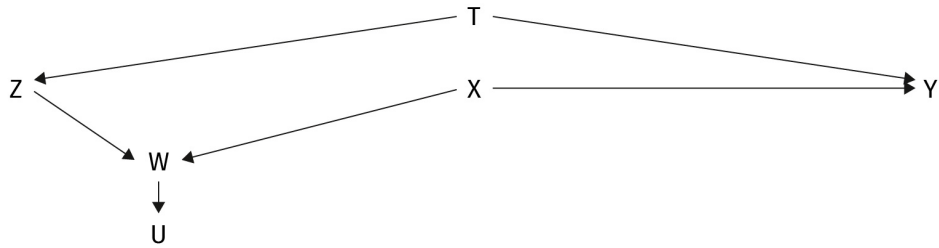
Here, we want to determine whether the nodes Z and Y are dseparated or if they can be d-separated. First, we note that there is a fork involving the node T, i.e. T is a common cause to both Z and Y, and, hence, a confounder for these nodes. If not conditioning on T, the path $Z \longleftarrow T \longrightarrow Y$ is open and Z and Y are d-connected. There is another path between Z and Y: $Z \longrightarrow W \longleftarrow X \dashrightarrow Y$. This path is blocked because W is a collider and, following the above rules, blocks the path. However, if we condition on W, the path is open again according to the rules. The nodes are d-connected, because there is at least one path connecting the nodes, and we therefore expect an association. The same applies if we condition on U, because U is a descendant of the collider W.

Because X is a fork just like T, we can block the path again if we need to condition on either W or U. We can summarize this as follows:

- If we leave the graph as it is, then Z and Y are d-connected, because of the fork at T.
- If we condition on T, then Z and Y are d-separated, because both the conditioned fork and W, a collider, block the path.
- If we condition on either W or U, then Z and Y are d-connected, irrespective of whether we condition on T.
- If we condition on T, then W and/or U, as well as X, and the nodes Z and Y are d-separated again, because X is the central element of the fork $W \longleftarrow X \longrightarrow Y$.

This example illustrates that analyzing causal graphs becomes quite complex even with a small number of variables. In particular, we need to pay close attention to the elements along a path connecting two nodes to determine whether any two nodes are d-connected or d-separated, i.e., if we can expect to find an association between the variables in the data. We also have to be careful which variables we condition on when analyzing the data.

Conditioning on the wrong variable, for example on a collider, will open a previously blocked path and we can expect a spurious correlation. On the other hand, we need to condition on confounders to block the path. The causal graph allows us to determine whether or not it is possible to block all paths and remove the expected association. This is important because any spurious association can lead to a bias in our analysis in general.

If two variables are d-separated without conditioning on any nodes along the path connecting them, these variables are unconditionally or marginally independent. If the variables become d-separated after some element along the path has been conditioned on, the variables are said to be conditionally independent, given a set of variables that has been conditioned on.

2.6 Conditional Independence

We have previously seen that we expect variables to be associated (or not) depending on how they are related to each other in a causal graph. We have also found that conditioning on variables can make them become dependent and we expect to find a correlation between them in the data - or, conversely, that conditioning on variables makes them independent.

Here, we want to give a more formal and thorough definition of the terms "independence" and "conditional independence." Two variables are independent if observing one does not influence the other.

For example, while having a cough increases the likelihood that you may be ill (where we make an observation that someone has a cough), noting that there are five books on your table has no impact on the probability.

Formally, we can express the independence for two events A and B as

$$P(A\,|\,B) = P(A) \tag{2.5}$$

This means that observing the value of the variable B does not give us any further information on the likelihood that event A occurs. If the above relation does not hold, then A and B are dependent. Dependence and independence are symmetric: if A is independent or dependent of B, then B is independent or dependent of A.

We can define dependence and independence for the variables X and Y in the same way as we defined it above for events A and B. Here we say that the variables are independent if for every value x of variable X and every value y of variable Y

$$P(X = x\,|\,Y = y) = P(X = x) \tag{2.6}$$

and, correspondingly, $P(Y = y\,|\,X = x) = P(Y = y)$, as, again, dependence and independence are symmetric. Variables are dependent if the above equation does not hold for at least one combination of pairs of values for the variables X and Y.

In addition to this absolute independence, two events (A and B) can be independent depending on a third variable (C). In this case, A and B are independent given C if

$$P(A\,|\,B,C) = P(A\,|\,C) \tag{2.7}$$

This means that in the presence of event C (or, if we condition on C), A and B become independent, i.e., the distribution $P(A|B,C)$ is independent of B. We can understand this intuitively by looking at the example of fire detectors (Pearl et al., 2016, p. 10): The event "detector is on" is dependent on "there is a fire." However, these detectors do detect the presence of smoke and not fire itself. If we now fix the value of the event C to "there is smoke" (i.e. condition on C and only look at events where there is smoke), we find that the detector is always on, regardless of whether there is a fire nearby.

We can express this condition for variables as well where we adopt a more formal approach.

📖 CONDITIONAL INDEPENDENCE

Let X, Y, and Z be variables and $P(\,\cdot\,)$ a probability distribution over some variables. Then, X and Y are conditionally independent given Z if (Pearl, 2009, p. 11):

$$P(x\,|\,y,z) = P(x\,|\,z) \quad \text{whenever} \quad P(y \wedge z) > 0$$
$$\forall\, x \in X, y \in Y \quad \text{and} \quad z \in Z. \tag{2.8}$$

We can express this in more detail: if for any combination where the variable X takes the value x, Y takes the value y and Z takes the value z and we have $P(Y = y \wedge Z = z) > 0$, then $P(X = x\,|\,Y = y \wedge Z = z) = P(X = x\,|\,Z = z)$. Informally, when we know that the value of Z is z and the probability distribution is greater than zero, $P(Y = y \wedge Z = z) > 0$, all information is already contained in Z. We do not learn anything else about X if we also know the value of Y,. In this case, knowing the value of Z is enough. We can say "z screens off X from Y" (Pearl, 2009, p. 11).

The symbol $\perp\!\!\!\perp$ is often used to indicate conditional independence (Dawid, 1979). Using this symbol, we can write Eqn. (2.8) as:

iff
"iff" is short for "if and only if."

$$(X \perp\!\!\!\perp Y\,|\,Z) \quad \textbf{iff} \quad P(x\,|\,y,z) = P(x\,|\,z). \tag{2.9}$$

Note that instead of using the symbol $\wedge$ for "and," a comma is often used as an abbreviation. Hence, the following notations are equivalent $P(A \,\text{and}\, B) = P(A \wedge B) = P(A, B)$.

📖 SUMMARY

In this unit, we have discussed how causal graphs can be used to analyze the relationship between variables. In particular, we have found that correlation or association between variables may be counterintuitive

---

and relying on correlations alone may lead to wrong results, as these correlations may be spurious. We have used directed acyclic graphs to represent causal relationships and we have learned how we can analyze such a graph to determine whether or not to expect an association between variables in the data. These graphs are built from nodes that are connected with arrows, with each node representing a variable. Special configurations of arrows and nodes such as chain, fork or collider determine the properties of the causal graph. The d-separation criterion allows us to determine whether we can expect an association between any two nodes or if we can remove a spurious association by conditioning on a variable along the path. Granger causality is a concept related to time series where one time series improves the prediction of another, for example, a specific feature causes the occurrence of a specific behavior at a later time in another time series.

HERE YOU CAN FIND THE VIDEOS FOR THIS UNIT.

Source: http://www.kaltura.com/tiny/odpjs

Alternative text:

UNIT 3

INTERVENTIONS

On completion of this unit, you will have learned ...

– the difference between observations and interventions.
– what confounders are and how to take them into account in causal analysis.
– what counterfactuals are and how to use them to explore the "world that would have been."
– what a randomized controlled trial is and why they work.
– when we cannot perform a randomized controlled trial.

Introduction/Case study

When we perform statistical studies and causal analyses, we are typically not content with describing the issue at hand. Instead, we want to take action in some way. For example, when developing a new medical drug, we want to establish how patients respond once they receive it. This is called an intervention, that is, actively do something and exploit a causal relationship, e.g., between administering the medicine and the health of the patient.

Everyone has probably heard the famous sentence: "correlation does not imply causation." To that effect, a correlation between variables does not necessarily indicate that using these variables to describe or predict a given effect is a valid and useful approach. On the contrary, we have previously discussed how spurious correlations can be created by conditioning on some variables or failing to condition on others, and, without access to the causal graph, it is very difficult to establish the causal effect of one variable on another.

In some scenarios, we can establish a causal effect using randomized controlled trials where all factors apart from one are either static or vary randomly. By influencing this one factor, we can establish a causal relationship.

For example, when testing a new drug, we can choose a group of patients at random who get the drug and a control group that does not.

If we then determine the outcome of the study, we can infer that the new drug works or it does not. However, in many cases, we cannot perform such randomized controlled trials. Under certain circumstances, performing the study may be unethical. For example, it would be unethical to force a large group of randomly selected individuals to heavily smoke over a long period of time simply to determine whether they would be more likely to develop and die of lung cancer than those who do not smoke. In other cases, it may be not be practical to perform a randomly controlled trial on a large scale, or it may not be possible to control external efforts such as the weather.

In this unit, we want to understand the difference between observing and intervening or "seeing versus doing." Central to this are interventions, i.e., the answer to the question "what happens if I do…" and counterfactuals, which describe an alternative reality: "if I had done X instead of Y, what would have happened?"

3.1 Seeing versus Doing

A core aspect of working with causal models is the understanding of how the system described would behave if we changed it, either by actually performing the change or by theorizing about what would have happened if we had done so.

This difference between observations and interventions (or "seeing versus doing") can be illustrated by the firing squad example (Pearl & Mackenzie, 2018, p. 39ff) illustrated in figure below.

**Figure 23: Causal Diagram for the Firing Squad Example.**
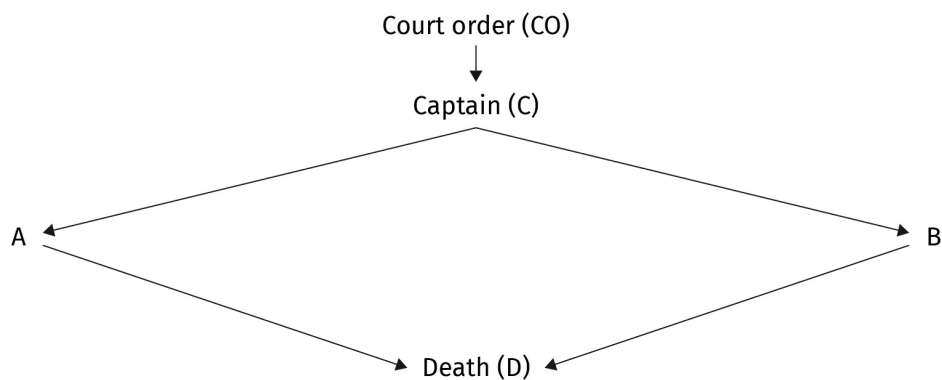
In this example, a prisoner is to be executed by firing squad. The squad is divided into two teams, A and B. The execution must be ordered by the courts (court order CO in the diagram), the order is passed to the leader of the squad (Captain C) who gives teams A and B the order to fire. As soon as the order is given, both A and B obey the order and fire, which results in the death D of the prisoner. Each of the variables (CO, C, A, B, D) is binary and can be represented by "true" (1) or "false" (0). Using the graph in figure entitled as "Causal Diagram for the Firing Squad Example", we can dissect what has happened. For example, if we observe that the prisoner is dead ($D = 1$), we can conclude that both team A (i.e., $A = 1$) and team B (i.e., $B = 1$) have fired, following an order given by the captain (i.e., $C = 1$), which will only occur if the court order was issued (i.e., $CO = 1$). Hence, we know that if the prisoner is dead, a corresponding court order was issued, because this is the only constellation in which the prisoner can end up dead. Suppose we observe that team A fired. What could we conclude about team B? Because of the causal structure, team A would only fire if the court order was issued, which, in turn, means that the captain would had to have given both teams the order to shoot. Hence, observing that team A shoots implies that team also B shoots, since both teams obey the order from the captain. Note that this reasoning holds even though A is not a cause of B, i.e., there is no arrow pointing from A to B.

We also observe that we cannot reconstruct this causal graph from observational data alone. If we are to record the variables, we only have two types of events: all variables are true or all variables are false. Hence, all variables are perfectly correlated, even though we know from our expert knowledge that A and B are only associated because they have a common cause (the captain's order) and don't share a causal relationship otherwise. We can find many analogous examples to illustrate that observational data alone do not explain our world. For example, without interventions, we would abstain from seeking medical expertise, as visiting a physician is strongly associated with being ill. We understand that someone going to a physician is correlated with being ill. However, that observation is not what makes them ill. On the contrary, that person goes to the physician to seek a (medical) intervention. Likewise, raw observation alone would suggest that firefighters might be related to fires erupting, as we only see them in times of emergency. From such observations, one could erroneously come to the conclusion that we should disband all fire brigades. However, we know that this is nonsensical, as the firefighters are only there to intervene.

We can now investigate what would happen in the firing squad scenario if we were to intervene. Imagine that team A always fires, regardless of whether the captain issues an order or not. This is not compatible with the causal graph we have discussed so far. Instead, we need to change the causal graph, as seen in part a of figure below.

Figure 24: Interventions in the Firing Squad Example.
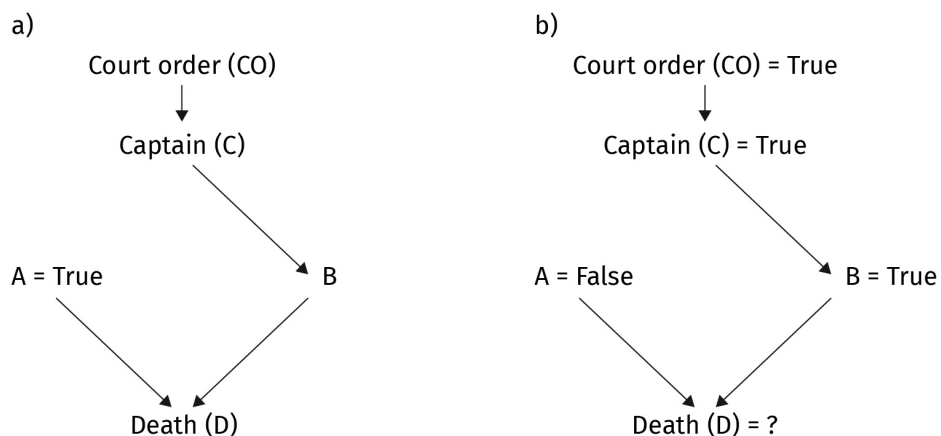
Since we intervene (i.e., set $A = 1$, regardless of any order the captain may have given), we force variable A to take a specific value and erase any arrows pointing to A. We do this because there is then no other cause that can influence the value of A and we force the variable to take a specific value. As we can see from the modified causal graph, the prisoner will always die. Furthermore, we also conclude that team B likely didn't shoot because B is still waiting for the command from the captain, who, in turn, needs a court order. Since team A always shoots, it's less likely that a court order was given.

Finally, we can investigate the **counterfactual** situation to the original situation from figure entitled as "Causal Diagram from the Firing Squad Example". Suppose the court order was issued, the captain gave the firing order, team B complied, but team A decided not to fire (despite the order). Would the prisoner be dead? This is called the counterfactual to the original situation, as we are taking a fictitious situation into consideration: Normally, team A would obey the order and fire—but what would happen if it did not? In this case, we remove the arrow from C into A to indicate that A does not obey the order ($C = 1$) and set $A = 0$.

Unfortunately, even though team A changes its course, the prisoner will still die, as team B will still carry out the order.

We can illustrate the concept of counterfactuals with another example: Suppose a patient follows a treatment and takes a specific medicine. We then observe a specific outcome, for example, the patient gets better. Then, we can investigate the counterfactual. What would have happened had the patient never taken the medicine? By definition, we cannot observe the actual outcome of the counterfactual scenario, as the patient has already taken the medicine and we have observed the corresponding outcome. To consider a counterfactual is to establish a fictitious world where we can go back in time, make sure that the patient does not take the medicine, and, crucially, change nothing else. If the patient does not get better in this hypothetical world, we would conclude that the medicine had a causal effect on the patient's real-world outcome.

Using these examples, we can also understand the difference between intervening on a variable and conditioning on them.

> 📖 **DIFFERENCE BETWEEN INTERVENTION AND CONDITIONING**
>
> The difference between intervening on a variable and conditioning on it is as follows (Pearl, 2009, p. 54): Intervening on a variable means that we fix the value of the variable and erase the arrows leading into the corresponding node in the graph. Hence, we change the system (and the graph). If we condition on a variable, however, we restrict the variable to a subset of values, but we change neither the real system nor the causal graph representing it.

It is important to note that interventions refer to individuals, whereas conditioning generally refers to populations. In particular, when we condition a variable X to the value x, we

observe the value y of variable Y with probability $P(Y = y \mid X = x)$. This means that $P(Y = y \mid X = x)$ describes the distribution of the variable Y for the case that $X = x$, i.e., for the subset of individuals in which the value of the variable X happens to be x. On the other hand, if we intervene, we force each individual in the population to take the value $X = x$ (Pearl et al., 2016, p. 55). To make this distinction explicit, we introduce the following notation:

> **📖 DO-OPERATOR**
>
> When we intervene on a variable (as opposed to conditioning on them), we express this as $do(X = x)$ (Pearl et al., 2016, p. 55).

Hence, $P(Y = y\,|\,X = x)$ is the probability that $Y = y$ conditional on finding $X = x$, whereas $P(Y = y\,|\,do(X = x))$ is the probability that $Y = y$ if we force $X = x$ through our intervention (Pearl et al., 2016, p. 55).

In the discussion above, we have implicitly assumed that the intervention is binary, as in the example of the firing squad. There, we considered the case where team A would fire irrespective of whether the captain (C) would issue the order. However, in general, interventions will follow a dynamic policy (Pearl et al., 2016, p. 70 ff). In these cases, the value of the variable X that we intervene on is specified by another variable or set of variables in a specific way. For example, we can imagine that the value x of variable X is given by $x = g(z)$, where $g(z)$ is some functional form that depends on another variable Z with value z. We can write this as $P(Y = y\,|\,do(X = g(z)))$.

As a concrete example, consider a physician treating a patient. The dose of the medicine the patient receives depends on the value of certain measured parameters such as blood pressure. If a patient's blood pressure is too high, they will be asked to take the medicine and the dose of that medicine depends on how high the blood pressure is. In this scenario, we say that the action (i.e., force $X = x$ for all individuals) is conditional on the value of variable Z.

3.2 Confounders and Counterfactuals

We encountered both confounders and counterfactuals when we investigated how to build causal graphs and how to distinguish the concepts between observation and intervention, i.e., seeing versus doing.

3.2.1. Confounders

Earlier, we saw that confounders generally rise when there is a common cause of multiple effects. For example, we have seen that both yellow fingers and lung cancer are caused by smoking. Looking at data, we expect to find an association or correlation between yellow fingers and lung cancer, even though yellow fingers are not a cause of lung cancer. In this example, smoking is a confounder. If we didn't include it in our analysis, we would include the variables "yellow fingers" and "lung cancer," as they are correlated.

However, the correlation is spurious. If we condition on smoking, e.g., only look at people who do not smoke, the correlation disappears; having yellow fingers does not alter your chance to contract lung cancer if you do not smoke. This also applies to variables that are

continuous and not binary. Pearl explains this with another example (Pearl & Mackenzie, 2018, p. 138): Imagine we test a new medicine and split the participants of the study into two groups, one that receives the drug and one that does not. However, it turns out that, on average, the participants in the group receiving the medicine is younger than those in the control group. Hence, the age of the participants becomes a confounder; we cannot directly translate the results from the study, as the two groups may behave differently due to the age of the participants. However, we can control for age and compare the two groups by stratifying by age, meaning that we form subgroups according to age within each group. We can then take the weighted average over all age groups, taking the relative population in the group into account, and then compare the group that receives the medicine to the control group.

While controlling for age is the right thing to do in the above scenario, it raises an important question: Which variables should we control for? Naïvely, the safest bet seems to be to control for everything. We could control for any variable imaginable, including age, gender, weight, height, etc. However, it is likely that we would only complicate the situation; we have already seen that conditioning on colliders introduces a (spurious) association between variables. Controlling for variables in the central element of a collider will make the results worse than if we leave the variables alone. Causal diagrams allow us to determine which variables we need to control for and which must not be controlled for.

Historically, there have been several ways to define confounding and confounders. One definition is: "A confounder is any variable that is correlated with both X and Y" (Pearl & Mackenzie, 2018, p. 152).

Another definition is given by Hernberg(Hernberg, 1996, p. 316): "Formally one can compare the crude relative risk and the relative risk resulting after adjustment for the potential confounder. A difference indicates confounding, and in that case one should use the adjusted risk estimate.

If there is no or a negligible difference, confounding is not an issue and the crude estimate is to be preferred. Personal judgment comes into play when what is "negligible" is decided. Some authors show both estimates and leave the decision to the reader." Informally, Hernberg suggests comparing the results when controlling and not controlling for a variable. If the difference between them is small, that variable is not a confounder and we can use the result when not controlling for that variable. As Hernberg points out, this approach leaves much to the interpretation of the author or the reader: which variables do we consider for controlling? Even in the best cases, we cannot look at all conceivable variables. Furthermore, what does "negligible" mean? The above discussion highlights that this definition of confounding is unlikely to result in a stringent approach.

Another approach to define confounding is the "classic epidemiological definition of confounding" that consists of three parts (Morabia, 2010)(Pearl & Mackenzie, 2018, p. 152): A confounder of X (treatment, e.g., medicine being administered) and Y (outcome, e.g., patient gets better) is a variable Z:

1. that is associated with X in the population at large.
2. that is associated with Y among people who have not been exposed to treatment X.
3. that should not be on a causal path between X and Y.

The third part of the definition is a relatively recent addition. Furthermore,

we should note that both the first and the second part of the definition do not require any causal links. The definition is based entirely on statistical concepts, and Z is assumed only to be correlated to X and Y but is not required to have any causal connection. However, consider the following constellation:

$$X \longrightarrow Z \longrightarrow Y$$

In this case, Z fulfills the "classical" definition above, i.e., the first two points. However, Z is not a confounder. Rather, it is a mediator, as it lies on a causal path between X and Y. Now consider the case where Z is a descendant of M in the chain $X \longrightarrow M \longrightarrow Y$, which we can illustrate as

$$X \longrightarrow M \longrightarrow Y$$
$$\downarrow$$
$$Z$$

In this case, Z is associated with X and Y and fulfills the first two requirements of the epidemiological definition (as before). However, now Z is a descendant of the mediator M and does not lie on a causal path between X and Y. Hence, the third requirement of the definition is also fulfilled, and yet, controlling for Z would be a disaster; since Z is a descendant of M and M is a mediator in the chain $X \longrightarrow M \longrightarrow Y$, Z acts as **a proxy** of M, and controlling for Z has the same effect as controlling for M, at least to some degree. As we have discussed before, proxies are generally not perfect substitutes for the variable, but controlling for them has (almost) the same effect as controlling for the "real" variable. For example, we might take the membership to a religious community as a proxy for religious beliefs or the membership in a political party as a proxy for political orientation.

Using the do-operator, we can define confounding more formally:
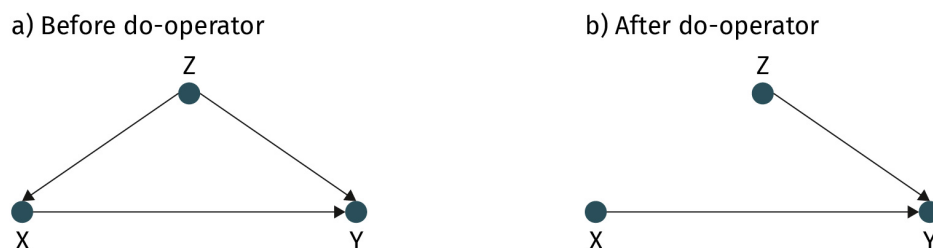
> **CONFOUNDING**
>
> Confounding is whenever $P(Y|X) \neq P(Y|do(X))$ (Pearl & Mackenzie, 2018, p. 151) (Pearl, 2009, p. 184).

Here, $P(Y|X)$ is the conditional probability that we observe some value of $Y = y$ given that we have observed $X = x$. The quantity $P(Y|do(X))$ describes the probability of observing $Y = y$ if we perform a (hypothetical) intervention that forces $X = x$. Whenever observing the variables taking a specific set of values for X and Y and forcing $X = x$ and observing the response of Y leads to a different result, we know that there is a confounder

that we haven't yet accounted for. Coming back to the example of how smoking is a common cause and confounder of both yellow fingers and cancer, we can observe the ratio of people with yellow fingers who develop cancer at some point in their life. However, if we paint everyone's finger yellow, that does not alter the chance of getting cancer. Therefore, there must be something that causes both cancer and yellow fingers, and we need to look for a confounder and avoid confounding.

We now investigate how we can calculate the causal effect in the presence of a confounder (Pearl et al., 2016, p. 55ff). Suppose X represents the treatment in a medical study, for example, getting the drug ($X = 1$) or not ($X = 0$). We then want to estimate if the new medicine has an effect, i.e., if the difference $P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0))$ is not zero when $Y = 1$ represents that the patients get better. This is also called the "average causal effect" (ACE). In this simple example, we only look at whether we administer the new drug at all. However, generally speaking, we could also investigate different doses and grades of improvement seen in the patients. Hence, X and Y may have several different values or be continuous. If there is no confounding, we can simply perform the trial and compare the results. However, we suspect that there is a third variable involved that may confound the results. For example, we may suspect that the gender of the patients plays a role. We can then introduce a confounder (Z) representing the gender and obtain the graph in part a of figure below.

Figure 25: Graphs for Adjustment Formula

a) Before do-operator

b) After do-operator

Z is a common cause to X and Y and is, therefore, a confounder—just like in the smoking example concerning yellow fingers and cancer. However, in our current example there is also a direct causal effect of X on Y, as we, of course, hope that the new medicine does indeed have a causal influence on the recovery of the patients. In order to proceed, we need to compute the probability $P(Y = y \mid do(X = x))$, i.e., the probability that we observe $Y = y$ when we force $X = x$. Forcing $X = x$, i.e., applying the do-operator modifies the causal graph, as shown in part b of in figure entitled as "Graphs for Adjustment Formula": Since we force the value of X, there cannot be any causal connection from the confounder Z to X. Any influence is severed, since we now control the value of X ourselves. Consequently, we remove the arrow from Z to X. In terms of probabilities, the conditional probability of the modified graph in part b of the figure is then the same as when applying the do-Operator: $P(Y = y \mid do(X = x)) = P_m(Y = y \mid X = x)$. A key observation is that Z is not influenced by our intervention as symbolized by the do-Operator. While we have removed the arrow from Z to X, the values of Z remain the same. In the medical

example, whether we make the patients take the medicine ($X = 1$) or not ($X = 0$) does not change their gender (Z). Hence, if we have a ratio of 50% to 50% male/female at the start, we will have the same ratio after applying the do-operator. Or, in the language of statistics, the marginal distribution of $P(z)$ remains invariant. Furthermore, the probability **P(Y=y|Z=z, X=x)** remains the same as we do not change the arrows from X into Y or from Z into Y. Informally, the probability remains unchanged, because the outcome in Y will not change even if we observe $X = x$ or force ($do(X = x)$). Hence, we observe the invariance conditions in the modified graph following the intervention using the do-operator:

$$P_m(Z = z) = P(Z = z) \tag{3.1}$$

$$P_m(Y = y \mid Z = z, X = x) = P(Y = y \mid Z = z, X = x) \tag{3.2}$$

We also note that X and Z are d-separated in the modified graph, as there is no connection between the variables. We have removed the arrow from Z to X, and Y is a collider on path X via Y to Z, blocking the path unless conditioned on (which we don't do). This implies that

$$P_m(Z = z \mid X = x) = P_m(Z = z) = P(Z = z) \tag{3.3}$$

In the first part of this equation, we make use of the fact that Z is independent of X and, in the second, that the intervention via the do-operator does not change the marginal distribution of Z. We can then compute the effect of the intervention:

$$P(Y = y \mid do(X = x)) = P_m(Y = y \mid X = x) \tag{3.4}$$

$$= \sum_z P_m(Y = y \mid X = x, Z = z) P_m(Z = z \mid X = x) \tag{3.5}$$

$$= \sum_z P_m(Y = y \mid X = x, Z = z) P_m(Z = z) \tag{3.6}$$

In this derivation, Eqn. (3.4) follows immediately from the above definition of applying the do-operator: This is how we arrived at the modified graph. In order to arrive at Eqn. (3.5), we make use of the **total law of probabilities** (Pearl et al., 2016, p. 13):

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n)$$
$$= \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$$

where the sum runs over all values that index i can take, i.e., all possible "sub-events" B_i that may contribute to A. The reason why we describe the probability for A like this is because it is often easier to describe the conditional probabilities of individual events and

the probability that these occur than the total probability that A will occur. In our case, the variable Z can take two values: male and female. We then exploit the fact that X and Z are independent (d-separated) to arrive at Eqn. (3.6). Using the invariance conditions in Eqn. (3.1), we can express the effect of the intervention using the do-operator.

📖 **ADJUSTMENT FORMULA**

$$P(Y = y \,|\, do(X = x)) = \sum_z P(Y = y \,|\, X = x, Z = z)P(Z = z)$$

(Pearl et al., 2016, p. 57)

This adjustment formula describes what we mean by "controlling for Z": We compute the association between the X (called the "treatment") and Y (the "outcome") for each possible value of the confounder Z and then take the average. It is important to note that this can be determined from the data as the right hand side of the adjustment formula only uses observational values. It is also important to note that, in this case, Z is always a parent of the treatment X and an observable confounder. This means that Z is a "real" variable we can measure and that the arrows emerging from Z point into X. The parent is often denoted pa (for parent) instead of Z in the adjustment formula, i.e., $P(Y = y \,|\, X = x, PA = z)P(PA = z)$ (Pearl et al., 2016, p. 59).

We can re-write the adjustment formula using the rule of product composition for DAGs (Pearl, 2009, p. 29): In **acyclic** graphs, the joint distribution of the variable is given by the product of the conditional probabilities from the parents to the children in the nodes:

Acyclic
In an acyclic graph, no path between nodes points back to a starting node.

$$P(x_1, x_2, ..., x_n) = \prod_i P(x_i \,|\, PA_i) \qquad (3.7)$$

where PA_i are the parents of the nodes x_i.

In our graph, variable Z is the parent of X and Y. If we multiply and divide the summand of the adjustment formula by $P(X = x \,|\, Z = z)$ (which doesn't change anything as it is equivalent to multiplying by one), then the numerator becomes $P(Y = y \,|\, X = x, Z = z)P(X = x \,|\, Z = z)P(Z = z)$, which is the joint distribution $P(X = x, Y = y, Z = z)$. The new adjustment formula is (Pearl et al., 2016, p. 59):

$$P(y \,|\, do(x)) = \sum_z \frac{P(X = x, Y = y, Z = z)}{P(X = x \,|\, Z = z)} \qquad (3.8)$$

where we remember that Z is a parent of X. The quantity $P(X = x \,|\, Z = z)$ is also called the "propensity score."

3.2.2. Counterfactuals

As we have seen earlier, counterfactuals explore the world that would have been. The interventions discussed above relate to a population or a group such as in the following scenario: what is the causal effect if we force all members of the medical study to take the medicine versus prohibit them from taking it? Conversely, counterfactuals predominantly apply to individuals.

For example, "Jon has taken the red pill and was cured of the disease—what would have happened had Jon taken the blue pill?" This is a purely hypothetical question, as Jon has already taken the red pill and we have observed the outcome. We cannot travel back in time and let Jon take the blue pill. We could, however, try to find someone similar to Jon and have that person take the blue pill. However, this ultimately only approximates an answer concerning Jon, as the two individuals are not identical.

The above discussion illustrates that the methods we have discussed so far are not sufficient and that we cannot express the counterfactuals with the do-operator. Instead, we need a new notation. Using the above example, we let X denote the treatment: Jon takes the medicine ($X = 1$) or does not ($X = 0$). Y denotes the outcome: Jon gets better ($Y = 1$) or does not ($Y = 0$). In the counterfactual world, we want to solve the following question: Given that we know that Jon took the medicine ($X = 1$) and got better ($Y = 1$), what is the probability that Jon's condition would have worsened ($Y = 0$) had he not taken the medicine ($X = 0$)? To express this, we use the following notation:

$$P\big(Y_{X=0} = 0 \,\big|\, X = 1, Y = 1\big) = ? \tag{3.9}$$

This notation highlights the difference between two different "worlds": We know the outcome of the "actual" world ($X = 1$) but would like to know the probability in a different world, one where $X = 0$. This difference is a critical aspect of counterfactuals (Pearl & Mackenzie, 2018, p. 287): If we didn't have hindsight into what has actually happened in the "real world," there would be no difference between $P\big(Y_{X=0} = 0\big)$ and $P\big(Y = 0 \,\big|\, do(X = 0)\big)$.

How do we then work with counterfactuals and determine the value we are interested in? Following Pearl (Pearl et al., 2016, p. 93 ff), we assume that we have some causal model M in which two variables, X and Y, are connected:

$$Y_x(u) = Y_{M_x}(u) \tag{3.10}$$

where Mx is a modified model in which we have replaced X with $X = x$, i.e, the counterfactual value we are interested in. Then, the counterfactual $Y_x(u)$ we want to compute is the solution to the modified model with the value set to the value $X = x$. So far, we have only considered graphical representations where the causal graph represents a statistical model. In order to work with counterfacutals, the original model M and the modified model Mx need to be a set of equations, also known as **structural causal equations**.

These can also be used to study the effect of interventions with the do-operator. However, the critical distinction is that we focus on individuals when considering counterfactuals, not populations.

We follow a simple example focused on students studying for a course (Pearl et al., 2016, p. 94 ff). The students are offered the chance to join a remedial program. This is the variable we can control, the "treatment" (X) where the value of X signifies the amount of extra tutoring time in the program. The students also do homework (H), and both the extra tutoring and the amount of homework a student does are causally connected to their performance during exams. Participating in the program can both help in the final score directly and when doing homework, which, in turn, helps improve exam performance. In our example, to avoid any selection bias, students are assigned to the program randomly. The resulting causal graph is shown in part a of figure below. The example makes a number of assumptions:

- All variables are standardized, i.e., they follow a Gaussian distribution with mean 0 and standard deviation 1. This means that if, e.g., the exam score is positive, the student scored better than average in the exam.
- The simple model assumes linear relationships between all variables.
  This implies we do not consider non-linear, higher order, or threshold effects.
- We assume that each variable is influenced at most by one unmeasured exogenous variable. These variables can have some influence on the variable in the graph, but we assume that these exogenous variables only influence one variable directly (and not two or more) and that there is no cross talk between these exogenous variables. In this simple linear model, these variables (U) represent the variation between students, i.e., some students will perform better than others, for example, due to talent.

Figure 26: Example for Counterfactual Reasoning
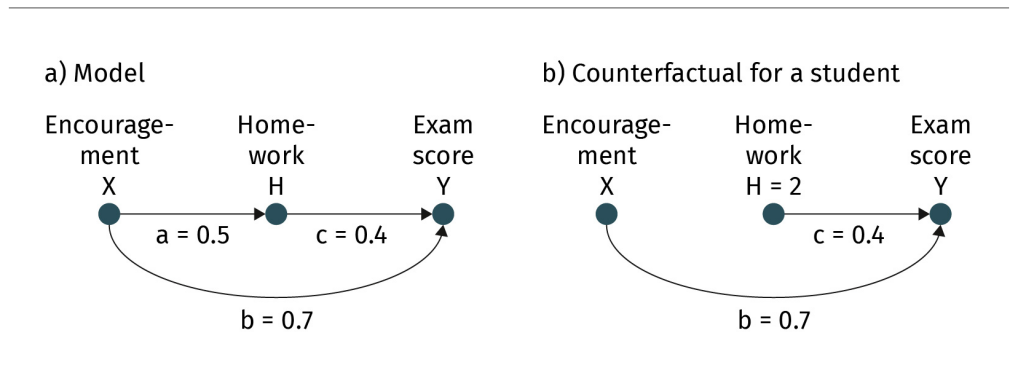
Then, the causal graph can be "translated," so to speak, into the following set of linear equations:

$$X = U_X \tag{3.11}$$

$$H = aX + U_H \tag{3.12}$$

$$Y = bX + cH + U_Y \tag{3.13}$$

$$\sigma_{U_i U_j} = 0 \quad \forall i, j \quad \in X, H, Y \tag{3.14}$$

This means that

- students are assigned randomly to the "treatment" X, meaning there is no arrow into X from external causes apart from the exogenous variable U_X that summarizes all unmeasured external influences.
  Hence, X only depends on U_X, as seen in Eqn. (3.11).
- the amount of homework a student does only "listens" to X, meaning there is an unknown coefficient a connecting the two and the exogenous variable U_H, as seen in Eqn. (3.12).
- the exam score is causally influenced both by the participation in the program and by the amount of homework, as well as an exogenous variable U_Y, as seen in Eqn. (3.13). In the linear model, these connections are represented by the coefficients b and c.
- all exogenous variables are uncorrelated and do not influence each other, as seen in Eqn. (3.14).

Before we can use the model, we need to determine the unknown coefficients (a, b, c). These can be measured using population data where we look at a large number of students and determine the numerical values. In the example, we assume that $a = 0.5$, $b = 0.7$, and $c = 0.4$ (Pearl et al., 2016, p. 95). With these values, we can now look at individual students. Suppose we measure for a particular student: $X = 0.5$, $H = 1$, and observe that this student scored better than the average student ($Y = 1.5$). This is the "real world," i.e, the one where we can measure both the treatment and the outcome. Using the equations above, we can use all values to determine the exogenous variables to be $U_X = 0.5$, $U_H = 0.75$, and $U_Y = 0.75$. As we have said above, these variables describe the "unique properties" of the student itself, i.e., their variation when measured against all other students.

We can then ask the counterfactual question: what would have happened had the student spent more time on homework instead of the time they have spent so far? What would have happened if $H = 2$ instead of $H = 1$, everything else being equal? To answer this question, we need to modify the model as shown in part b of in figure entitled as "Example for Counterfactual Reasoning". We remove the arrow from X to H, as the participation in the study program no longer has an influence on the amount of homework the student does. Instead, we set $H = 2$, since this is what we want to know. According to Eqn. (3.10), the value of the counterfactual is given by the solution to the modified model, i.e., the one in part b of figure entitled as "Example for Counterfactual Reasoning". We are interested in the counterfactual solution for the hypothetical "outcome" $Y_{H=2}$, i.e., the performance in the exam had the student studied signifianctly more than before, all else being remaining the same. Using the numerical values $b = 0.7$, $c = 0.4$, $U_Y = 0.75$, and $X = 0.5$ (the student still participates in the remedial program for the same time as before—the only thing we change is the amount of homework) and $H = 2$, we obtain $Y_{H=2} = 1.9$, i.e., the performance of the student is now almost 2 standard deviations above the average, up from $Y = 1.5$ in the "real world."

The same approach can then be followed for non-deterministic models.

However, in this case, we cannot uniquely identify the exogenous variables.

We need to assign a suitable probability distribution for each.

A related concept to the approach analyzing counterfactuals described above is that of "potential outcomes." This was developed by Rubin (Rubin, 1974). The framework uses the same notation to denote the counterfactuals: $Y_{X=x}(u)$ or $Y_x(u)$ is the counterfactual outcome for some "unit" (or individual) u if the value of the "treatment" we can control had been $X=x$. In fact, Pearl has taken the notation from the potential outcome model and used it for his analysis of counterfactuals (Pearl, 2009, p. 243). The main difference between Rubin's potential outcome framework and Pearl's counterfactual framework is that Pearl's framework is based on causal graphs that are connected to a structural model. The dependencies of the variables can be derived from the graph and causal model. In Rubin's framework, however, there is no underlying causal graph or structural model. Instead, the questions about counterfactuals are formulated algebraically. This leads to three assumptions that have to be accepted to work in that framework (Pearl & Mackenzie, 2018, p. 280)(Pearl, 2009, p. 100):

- The effect of a treatment on an individual is independent of what treatment (if any) the other individuals get. This assumption is generally fulfilled unless the treatment is a scarce resource in emergency situations. This assumption is called the stable unit treatment value assumption (SUTVA).
- The treatment is assumed to be "consistent," meaning that if you receive the treatment (e.g., take medicine), the effect remains the same regardless of whether you took part in the study. For example, you might take an aspirin against your headache. The headache would go away if you took the aspirin, as part of the study or in everyday situations.
- The variables must meet the requirement for conditional independence $Y(x) \perp\!\!\!\perp X \mid Z$ ("conditional ignorability").

The last assumption specifying the "conditional ignorablity" is the most difficult to understand. It can be interpreted in the following way: "The way an individual with attributes Z would react to treatment $X=x$ is independent of the treatment actually received by that individual" (Pearl, 2009, p. 100). This means that if we control for any confounders in Z, those individuals that have one potential outcome $Y_X=y$ are as likely assigned to either the treatment group (i.e. they receive the treatment $X=x$) or the control group (that does not receive the treatment), as individuals that have a different potential outcome $Y_X=y'$: The value of the potential outcome does not influence whether an individual would end up receiving the treatment. This is similar to the concept of "exchangeability" by Greenland and Robins (Greenland & Robins, 1986). This means that we randomly assign the participants to two groups: group A and group B.

Since the participants are assigned randomly, the two groups are homogeneous and have the same characteristics. Otherwise, we could not establish a causal effect from the observation of the outcome of the two groups, e.g., if one group in a medical study was healthy and the other not. We then make a choice, e.g., group A receives the treatment and group B doesn't and we observe the outcome. However, we could also have made the choice that group B receives the treatment and A doesn't. Hence, the groups are exchangeable. The challenge with the concept of "ignorability" is that it is difficult to confirm that it is

fulfilled. In the structural approach followed by Pearl, confounders are defined in the context of a causal graph with an associated model. Using this model, we can determine all confounders and determine the relationship between all variables. Admittedly, our model may be incomplete or wrong. Regardless we do have all (technical) means to develop and test such a model. The potential outcome framework by Rubin is not based on causal graphs. It is, therefore difficult to determine whether the assumptions, in particular the ignorability requirement, are fulfilled.

3.3 Causal Inference versus Randomized Controlled Trials

Thus far, we have encountered the difference between seeing and doing, where we compared observations and interventions, explored the "world that would have been" with counterfactuals, and removed confounding by adjusting for a specific set of variables for which we we have learned to be careful to distinguish between the variables that we should control for and those we do not. This should give us all the tools we need to estimate the causal effects. However, when looking at the literature, the randomized controlled trial (RCT) is seen as the gold standard of establishing causal effects. For example, if a new medicine is to be approved, it first needs to be tested in a clinical trial. The RCT was popularized by R. A. Fisher (Box, 1978, chap. 6), who used the example of a field that is to be treated with one of two possible options for fertilizers. To find out which fertilizer is better, Fisher discusses an experimental approach. The farmer could use fertilizer A on one half of the field and B on the other. However, this would be subject to confounders, as the two halves may be different. The top half may have a different drainage than the bottom half, or the left half a different texture or intrinsic fertility than the right. Swapping the halves in the next year would introduce weather as a confounder.

Initially, Fisher then devised an elaborate grid called "Latin square" to cover all combinations of fertilizer, soil type, plants, etc. Yet, however elaborate such a testing scheme is devised, one can always think of one more confounder that needs to be included. In the end, Fisher realized that only random assignment would solve this problem; the experiment needs to be repeated many times to account for natural (statistical) variations, and the type of fertilizer is assigned randomly. This is the random controlled trial because we are assigning the "treatment" randomly within a controlled study. In a clinical setting, we would take all the test subjects and randomly assign them into group A and group B. We then decide (randomly) whether group A receives the medicine we want to test and B the **placebo** (or the other way round). In this case, the random aspect defines in which group each individual is placed. There is no mechanism other than a random number influencing how the decision of which participants receive the medicine is made. To avoid any lurking bias, the trial is typically performed **double-blind**. The crucial aspect is that this randomization erases all arrows pointing from potential confounders into the variable (X) describing the treatment. Because the treatment is assigned randomly, there is no possibility that this assignment can be influenced by any other variable. Hence, even if there are confounders present, they are removed. This is why randomized controlled trials work and

Placebo
In medicine, a placebo is a substance that looks like medicine but has no effect.

Double-blind
In a double-blind study, neither the patient nor doctor know if the patient receives the actual medicine or a placebo.

allow us to establish the causal effect. We can, of course, only observe the actual outcome in the groups of the trial, e.g. "received treatment" ($Y_{X=1}$) and "has not received treatment"($Y_{X=0}$), i.e., the treatment and the control group.

Randomization ensures that the means of the outcomes in the study will converge to the means if we could observe all potential outcomes. The causal effect observed in the trial will hence converge to the "true" causal effect $E[Y_{X=1} - Y_{X=0}]$ that we could determine if we could calculate all counterfactuals. This is because the randomization implies that there is no intrinsic bias when assigning the individuals to a group. This also explains why the treatment and control group need to be "sufficiently" large.

If the groups are too small, the statistical fluctuation of the observed effect may be too large compared to the causal effect we wish to establish. The smaller the effect, the larger the two groups need to be to get an accurate estimate of $E[Y_{X=1} - Y_{X=0}]$. This difference is also called the average treatment effect (ATE).

Remember that the causal effect is defined at the level of individuals. What we are really interested in is the causal effect on an individual. In such a scenario we observe the outcomes when the same individual has received the treatment ($Y_{X=1}$) or not ($Y_{X=0}$). This can, of course, not be observed in practice, since we cannot both give and not give the treatment to the same person. Hence, one of the outcomes is a counterfactual describing what would have happened. What we can observe is the average treatment effect ATE= $E[Y_{X=1} - Y_{X=0}]$ of those assigned to receive the treatment (or not) at the group level. This is because the expectation value is a linear operator and, hence, $E[X \pm Y] = E[X] \pm E[Y]$. In our case, this means that the average of the difference is equal to the difference of the averages, and, hence, we can write ATE = $E[Y_{X=1}] - E[Y_{X=0}]$. In a random controlled trial, we assign individuals randomly to the treatment and control group and, by observing the average in each group, we can estimate the average treatment effect.

Related to the ATE are the Average Treatment effect on the Treated (ATT) and Average Treatment effect on the Control (ATC). These refer to the average causal effect when looking only at the group that has received the treatment (ATT) or the control group (ATC). Note that both ATT and ATC contain an unobservable counterfactual, as we cannot measure the outcome of individuals in the treatment (or control) group that have not received (or have received) the treatment. In the case of a randomized controlled trial, ATE is the same as ATT, since we assume that both the group receiving the treatment and the control group have the same properties.

Whether we see an RCT as a gold standard is probably more a question of preference and of the social dynamics within a given research community.

The more important question is whether they are necessary given all we know about how to deal with confounders. The answer is: no, not really.

If we know all relevant confounders and measure them, we can adjust for them (Pearl & Mackenzie, 2018, p. 149). However, the randomization in the RCT ensures that the "treatment" is assigned randomly. This means that all arrows from any confounders pointing

into the treatment variable are severed, not just the ones we think of. If we do not perform a randomized controlled trial, we have to convince ourselves (and others!) that we have indeed considered and adjusted for all confounders. By design, an RCT does this for us automatically. Moreover, some confounders may be difficult to determine.

However, when we want to perform a random controlled trial, we have to make sure prerequisites and assumptions are fulfilled. The group of people participating in the trial needs to be representative of the population we want to analyze. For example, if we want to study the effect of extra tutoring on university students, looking at kindergarten children will not be helpful. We also have to look at how we obtain the participants of the study should the study involve individual persons. In most cases, we cannot grab people at random off the street and add them to the trial.

Typically, we need to work with volunteers. However, the act of volunteering may also introduce a bias. Some terminally ill patients may opt to participate in a study because they have nothing to lose. Their health is also severely compromised, which may skew their response to treatment.

In some studies, financial compensation is offered for those who volunteer, which again may introduce a bias regarding the people who are attracted by this, etc.

If the study runs for a prolonged length of time, we also need to make sure that the participants do not drop out of the study while the study is being conducted. For example, if we want to test the effectiveness of a new medicine targeting high blood pressure or cholesterol, the study will likely run for weeks or months. If the patients do not report regularly, the resulting data may be biased, particularly, if the patients do not drop out randomly. This is called "loss of follow-up." Typically, all aspects of the trial are defined before it is started. This includes, for example, the length of the trial, the actual treatment, and how it is administered. For example, it may be decided that a procedure is performed as the treatment or a medicine is administered in specific doses.

Ideally, one should then acquire the data in the study but not perform the full analysis until all data are recorded. This is done to avoid a potential bias that arises if an intermediate result we obtain while recording the data and analyzing the data while the study is performed, may show a large statistical fluctuation which in turn may mean that such an intermediate result may not be representative of the final result. However, that may lead to a conundrum, especially in medical studies: would it be ethical to withhold a treatment from the control group if the group receiving the treatment has already shown significant improvement at a preliminary analysis? Therefore, in many cases, data are analyzed at fixed intervals and a decision is made to continue or conclude the study. In a case of significant positive results, studies may then indeed be ended early and the treatment offered to all participants, e.g. Auvert et al. (2005); Gray et al. (2007); Bailey et al. (2007).

While RCTs do indeed offer practical benefits, performing one may not always be possible. For example, interventions may not be possible, or it may be unethical to force the participants to receive a treatment that we want to prove is harmful. In these cases, using causal analysis on observational data is the only way we can establish the causal structure and the causal effects of the aspect we are interested in.

📖 SUMMARY

One of the most critical aspects when understanding causal effects is the distinction between observations and interventions, i.e., "seeing versus doing." In observations, we see what happens in a given circumstance.

However, in most cases, we want to determine the effects of an intervention. For example, does this new medicine have any effect? What would happen if we did this or that? When we try to measure causal effects, we need to be sure that we take all relevant confounders into account. Confounders are variables that are, for example, a common cause for treatment and outcome and they can lead to wrong results if we do not account for them. Note that we are using here the terms "treatment" and "outcome." This is because causal inference has been used for a long time in epidemiology, therefore these terms tend to be used even if we are not considering medical or epidemiological examples. A treatment is understood to be an intervention we perform and the outcome is what we observe.

Counterfactuals are a powerful way to interrogate causal relationships.

They address the following question: "What would have been had I done x?" Counterfactuals are always hypothetical questions, as we have already performed an intervention and observed the result.

In many disciplines, randomized controlled trials are considered the gold standard for measuring causal effects. These trials work because the randomization procedure severs the effect of possible confounders.

However, there are a number of circumstances in which such a trial cannot be performed, such as forcing people to smoke to determine whether smoking is harmful.

HERE YOU CAN FIND THE VIDEOS FOR THIS UNIT.

Source: http://www.kaltura.com/tiny/r61oj

Alternative text:

UNIT 4

DO-CALCULUS

4. DO-CALCULUS

Introduction/Case study

We have previously encountered interventions (e.g., confounders) and elements that we can use to express the behavior of variables in directed acyclic graphs (DAG). Taking confounders into account, that is, to "adjust" for them (in causal analysis terms), is one of the most important aspects of establishing a causal effect from data. To do so, we first need to identify which variables act as confounders and then determine how these can be taken into account.

Thus far, we have primarily built an intuitive understanding of confounders and how to adjust for them. In this unit, we want to formalize the approach.

Additionally, establishing the relevant rules allows us to analyze more complex graphs than the ones we have seen in the earlier examples.

We have also encountered the do-operator already, which we use to express interventions such as $do(X = 1)$, where we force the value of the variable X to take the value of 1. For example, let's imagine that $X = 1$ refers to administering a certain medicine we want to test to a group of people and $X = 0$ to withholding it. Or, we could refer to different values of X describing a specific dose we wish to administer (or something else). In this unit, we will give a more detailed description of do-calculus that allows us to use the do-operator to establish causal effects. The main idea is that we need to transform expressions that contain the do-operator into others that can be estimated from observational data—only then are we able to use data outside a randomized controlled trial to infer causal effects.

4.1 Front- and Back-door Criterion

4.1.1. Back-Door and Front-Door Paths

We have previously seen how variables can be associated, even if there is no apparent (causal) relationship between them (e.g., yellow fingers and lung cancer being related even though yellow fingers are not causally connected to lung cancer). This means that if we painted fingers yellow, i.e., $do(yellow \quad finger)$, we would not affect the risk of getting cancer. However, both are associated in the data, since smoking is a common cause of either of them. In this case, smoking is a confounder, as shown in part a of figure below.

Figure 27: Common Cause
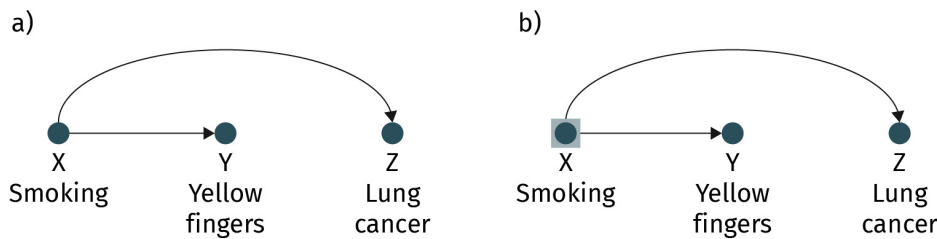
We then saw that we can remove this spurious association by conditioning on the confounder, i.e., by looking at the values of the variables "yellow fingers" and "lung cancer" for smokers and non-smokers separately. This is shown in part b of figure above, where the box around X indicates that we adjust for the confounder, i.e., look at specific values.As illustrated by this example, variables can be associated if there is a **path** between them. For example, in figure above we have the following paths: X to Y (in the direction of the arrow); X to Z (in the direction of the arrow); and Z via X (against the direction of the arrow) to Y (in the direction of the arrow).

We can formalize the different characterisations of the paths in the following way:

> **FRONT- AND BACK-DOOR PATH**
>
> - A front-door path is a causal path in the direction of the arrow between any two nodes X and Y in the graph.
> - A back-door path is any path between any two nodes X and Y that starts with an arrow pointing into X, i.e., against the causal direction (Pearl & Mackenzie, 2018, p. 158).

The front-door (or causal) paths represent the causal relationships we want to explore or we know are true. Informally, these are the "real" associations or correlations between variables that have a "deeper meaning." By this, we mean that they can explain what we observe, for example, that smoking is a cause of lung cancer. If we look at the population of smokers and nonsmokers, we will find that the population of smokers is more prone to lung cancer than the non-smokers and that smoking is the cause of lung cancer.

The back-door paths, on the other hand, are those that introduce spurious correlations or associations between variables in the data. In the example of smoking, we have a spurious correlation between "yellow fingers" and "lung cancer." We can start a back-door path between these variables by starting at Y with an arrow pointing into Y from X and then go to Z or, conversely, by starting with an arrow pointing into Z from X and then going to Y. Note that, according to the definition a back-door paths with an arrow pointing into

Path
A path is any connection between any two nodes (such as X or Y) in a directed acyclic graph (DAG), running either in the direction of the arrows or against the arrows.

one of the variables, it is not required that we go against the direction of the arrow along the entire path. In our example, we start with the first step against the direction of the arrow (hence making it a back-door path) and then move in the direction of the arrow for either variable X or Y.

In order to avoid spurious correlations, we need to treat the variables we have (or we can add) so that they block the path according to their properties.

This can be done using the following rules (Pearl et al., 2016, p. 46):

📖 PATH BLOCKING RULES

- A path can be blocked when conditioning on a fork or chain.
- A blocked path is opened when conditioning on a collider.

The confounder discussed above is represented by a fork in the DAG.

Hence, conditioning on the variable "smoking" blocks the back-door path, meaning we can establish the causal effect of smoking on lung cancer (or yellow fingers). Note that the same holds true if we condition on descendants of these elements. For example, we may open a blocked path if we condition on a descendant (or child) of a collider. We need to be careful when analyzing more complex graphs that we do not accidentally open paths we need to block by conditioning on children further down in the graph.

We can illustrate this at the slightly more complicated DAG shown in figure below. Note that we have already encountered this DAG earlier (Pearl et al., 2016, p. 48). In this example, we want to figure out if there are any spurious connections between Z and Y and if we can remove them— in other words, if there are any open back-door paths that we can close.

Against the arrow
A path against the direction of the arrow is an back-door path.

There are two paths with arrows pointing into Y. The first one is from Y to T (**against the arrow**), and then from T to Z (in the direction of the arrow). The node T is both a fork and a confounder because it is a common cause to both Z and Y. Assuming we can measure T, we can block this path by conditioning on T. Then, there is another path starting at Y going against the direction of the arrow (thus making it a back-door path again) to X from X to W (in the direction of the arrow) and from W to Z (against the direction of the arrow). Therefore, we have two back-door paths connecting Y and Z. However, that path is blocked because W is a collider and colliders block the path unless conditioned on. We therefore only need to condition on T to remove the spurious association of Z and Y. In this case, Z and Y are also said to be d-separated since there are no open back-door paths.

Figure 28: A Complex Causal Graph
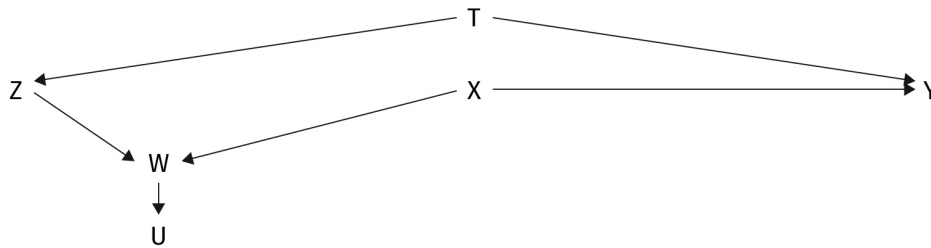
However, if we were to condition on W or on U (as a descendant of W), we would open the blocked path and Z and Y would become d-connected and associated again, even if the path via T were still blocked by conditioning on T. This might happen for two reasons: for example, we might make a mistake, which can easily happen if the graphs become more complicated.

Alternatively, we may want to measure the causal effect depending on W.

For example, we might want to know what the causal effect is for specific values of W. In this case, we need to condition on W to look at specific values. Another reason might be that we have no choice: in order to block some other back-door path crossing through that node, we need to condition on it. Remember that the function of a node is path specific: A node may be a collider on one path but a fork or chain in another. If we are forced to block that other path because the node is, for example, a fork there, we need to condition on it—even if that opens the path on which that node acts as a collider. We then need another way of closing the path again. In the example in figure above, we can also condition on X, which is a fork on the path between Z and Y. In this case, the path is blocked again.

Hence, we can d-separate Z and Y by either conditioning on T alone, T, W, and X, or T, U (because U is a descendant of W and X). We could also condition on T and X. However, the path through X is already blocked because of the collider in W, so we do not have to do this. Note that it would not do any harm.

4.1.2. Back-Door Criterion

We can now formalize our treatment and give a precise definition that describes the possibility of closing back-door paths. This is known as the "back-door criterion" (Pearl, 1993).

Let G be a given causal diagram with a set of variables (V) measured from **observational data**. We wish to establish the causal effect of the Both variables X and Y are part of the set of variables V, i.e., measured from data. Using the backdoor criterion we want to determine whether there is a subset of variables (Z) from V, i.e., $Z \subseteq V$ that we can use to block all back-door paths, thus allowing us to estimate the causal effect from the observational data.

Observational data
The term observational data refers to data obtained by observing a system without influencing it (as opposed to data from a controlled trial).

Although this definition sounds quite intimidating, it really means the same thing as our evaluation of the back-door paths in the example above—it is just more formal and applicable to any DAG G. Informally, we can summarize this as (Pearl et al., 2016, p. 61):

- We block all back-door paths, for example, by conditioning on forks or chains.
- We make sure that the **directed paths** in the direction of the arrows we wish to investigate are still open.
- We make sure that we don't accidentally open another back-door path that leads to spurious correlations in the data by controlling (or not) on the wrong element in the graph.

Directed paths
Directed paths in the direction of the arrow are also called "causal paths."

If the back-door criterion is fulfilled, we can estimate the causal effect of intervention (X), i.e., $do(X = x)$ on outcome Y using the back-door adjustment formula:

📖 **BACK-DOOR ADJUSTMENT**

$$P(Y = y \mid do(X = x)) = \sum_z P(Y = y \mid X = x, Z = z)P(Z = z)$$

(Pearl et al., 2016, p. 61)

A proof can be found in (Pearl, 2009, p. 80). Note that this is very similar to the adjustment formula we encountered earlier. The previous adjustment formula was specifically aimed at confounders that are direct parents of the intervention. The above back-door adjustment formula is more general and contains this case automatically.

4.1.3. Front-Door Criterion

The back-door criterion allows us to identify backdoor paths that lead to spurious associations in the data and identify confounders we need to control for. The back-door adjustment formula then allows us to estimate the causal effect from observational data. Unfortunately, in some cases, this approach will not work.

An example is shown in part a of figure below. The situation is similar to what we have already encountered: the variable U is a common cause to X and Y and is, hence, a confounder on a back-door path from Y to X.

Figure 29: Front-Door Criterion
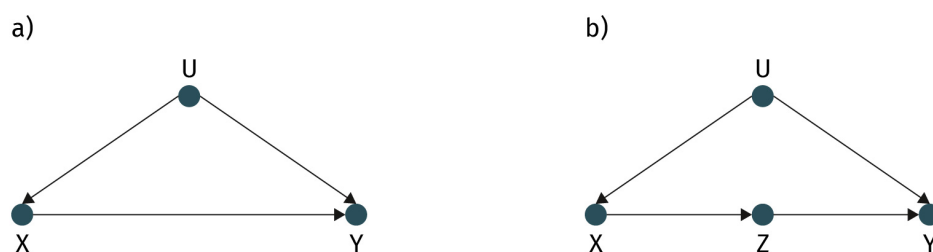
To block the path, we would need to condition on U—this is the approach we have taken so far. In these situations, we have assumed that we have data regarding this variable so we can condition on it. In the example of yellow fingers and lung cancer above, we would condition on "smoking" and look at the association between yellow fingers and lung cancer separately for smokers and non-smokers (or heavy and light smokers or smokers who smoke one, two, three, etc. cigarettes per day).

However, what happens if we do not have data on U, i.e., U is unobserved? In this case, we cannot close the back-door path and remove the spurious association. An example is the causal connection between smoking and cancer. In this example, X is smoking and Y is cancer. We want to establish the causal relationship and establish whether smoking indeed causes cancer. At the time, there was a major discussion, and the tobacco industry argued that the association between cancer and smoking is explained by a supposed "smoking gene," see, e.g., (Spirtes, 2000, p. 239ff).

In this argument, the gene takes the role of the unobserved variable (U), as we cannot measure the gene directly (at least, we could not at the time).

Later, it was discovered that there truly is a gene related to smoking (Lassi et al., 2016). However, the act of smoking still causes cancer.

Hence, in our example, U is unobservable, which means we cannot block the back-door path. For a long time, the tobacco industry argued successfully that the causal influence could not be proven.

Nevertheless, the causal effect of X on Y can be established under some circumstances. If we can identify a mediator that transports the causal effect from X to Y, we can determine the causal effect, even in the presence of unobservable confounders. This is shown in part b of figure entitled as "Front-Door Criterion". In this case, Z is a mechanism of the causal effect. In the example of smoking, Z is the tar deposits in the lung. Hence the causal chain is as follows: smoking leads (X) to tar deposits in the lung (Z) that cause cancer (Y).

We still cannot block the back-door path because U is still unmeasured, but we can exploit the new variable (Z) as mediator (Pearl et al., 2016, p.

68). The causal effect from X to Z can be identified immediately, as there is no back-door path from X to Z. Hence,

$$P(Z = z \,|\, do(X = x)) = P(Z = z \,|\, X = x) \tag{4.1}$$

which means that the observation is the same as the intervention. We can also identify the causal effect of Z on Y. There is a back-door path from Z to X (against the arrow, making it a back-door path), from X to U, and from U to Y. However, X is a **non-collider** on this path and according to the rules, we can block it by conditioning on X. We can do this because X is observable and we have data for this variable—indeed, this is the variable we wanted to analyze in the first place. Using the adjustment formula, we can then write

$$P(Y = y \,|\, do(Z = z)) = \sum_{x} P(Y = y \,|\, Z = z, X = x) P(X = x) \tag{4.2}$$

We now need to combine both effects, since we are interested in $P(Y = y \,|\, do(X = x))$, i.e., the result of the outcome when we perform an intervention on X. The idea is as follows: we do not intervene on Z directly as this mechanism. In the example of smoking, we do not add tar deposits into the lung ourselves. This is what happens due to the properties of the system we want to analyze. Hence, if the system "chooses" to assign the value z to Z, the probability of observing Y is $P(Y = y \,|\, do(Z = z))$.

However, since we perform the intervention $do(X = x)$, the probability of this is $P(Z = z \,|\, do(X = x))$. Taking all possible values of the mediator Z into account, we can combine the parts:

$$P(Y = y \,|\, do(X = x)) = \sum_{z} P(Y = y \,|\, do(Z = z)) P(Z = z \,|\, do(X = x)) \tag{4.3}$$

We can then use the expressions 4.1 and 4.2 to transform the right-hand side into do-free expressions that can be estimated from observational data:

$$P(Y = y \,|\, do(X = x)) =$$

(4.4)

$$\sum_z \sum_{x'} P(Y = y \,|\, Z = z, X = x') P(X = x') P(Z = z \,|\, X = x)$$

This is known as the "front-door" criterion and adjustment, which we can define more formally in the following way (Pearl, 1995):

FRONT-DOOR CRITERION

A set of variables (Z) satisfies the front-door criterion for an ordered pair of variables (X, Y) if the following conditions are met (Pearl et al., 2016, p. 69):

- Z intercepts all directed paths from X to Y.
- There is no unblocked back-door path from X to Z.
- All back-door paths from Z to Y are blocked by X

Essentially, Z is a mediator on all possible paths from X to Y, and we can establish the causal effects from X to Z and from Z to Y.

If the front-door criterion is fulfilled, we can establish the causal effect from observational data via the front-door adjustment formula.

FRONT-DOOR ADJUSTMENT

$$P(y \,|\, do(x)) = \sum_z P(z \,|\, x) \sum_{x'} P(y \,|\, x', z) P(x')$$

(Pearl et al., 2016, p. 69)

Note that, in general, we do not want to introduce a mediator, as we are typically interested in the total causal effect of an intervention, not just the one that is related to a specific mechanism expressed by the mediator.

Furthermore, if the mediator we choose is not the right mechanism to transport the causal effect, our conclusions will also be wrong. This is illustrated by the following example (Pearl & Mackenzie, 2018, p. 302ff): In the early days of long distance travels on the seas, scurvy was a dangerous disease for the sailors. It was observed that consuming citrus fruits eliminated the risk of falling ill to this disease. Soon after, all ships carried a supply of citrus fruit. It was therefore unexpected that, about a century after this problem was thought to be solved, expeditions to the polar regions were again plagued by scurvy. It was thought—but not proven—that citrus fruit prevented scurvy by virtue of their acidity, i.e., acidity was the mechanism by which the disease was prevented: citrus fruit $\longrightarrow$ acidity $\longrightarrow$ scurvy. However, a detailed analysis showed that it was vitamin C (and not any acid) that prevented scurvy. Therefore, the correct causal path is citrus fruit $\longrightarrow$ vitamin C $\longrightarrow$ scurvy.

4.2   The Three Rules of Do-Calculus

We have seen previously how the do-operator can be used to formalize interventions and derive the causal effect. The general idea behind using the do-operator is that we want to extract the causal relationships between the intervention or treatment and the outcome using observational data.

In the front-door and back-door adjustment formula, we have seen how we can make use of the special structure of these constructs to transform the expressions that contain the do-operator into those that do not. This is because we cannot observe probability distributions that contain the do-operator, but only those without, as these relate to observational data we can record.

The do-calculus (Pearl, 1995) provides three rules that are sufficient to transform to expressions that contain the do-operator into those that do not. However, this requires that that the causal effect is "identifiable". We call a causal effect identifiable if we have a causal graph G and we can use a finite number of transformations according to the rule of do-calculus that translate the expressions containing the do-operator into those that do not.

The latter can be determined from observational data (Pearl, 2009, p. 86).

Before we focus on the rules of do-calculus, we need to introduce further notation that relates to the various operations we can perform on a causal graph G, specifically, the removal of arrows emerging from or pointing into some node. We place a line over the variable, if we delete any arrows that point into some node. For example, if we start with the full causal graph G as shown in part a of figure entitled as "Graphs demonstrating various applications of the Do-Calculus Rules", part b shows the graph if we remove the arrow between X and Z. Since the arrow we have removed points into Z, the new graph after this operation is called $G_{\overline{Z}}$. Similarly, we use a line under the variable if we remove an arrow that emerges from the corresponding node. Since the arrow we removed emerges from node X, the same example (part b of figure below) can also be denoted as $G_{\underline{X}}$.

Figure 30: Graphs Demonstrating Various Applications of the Do-Calculus Rules
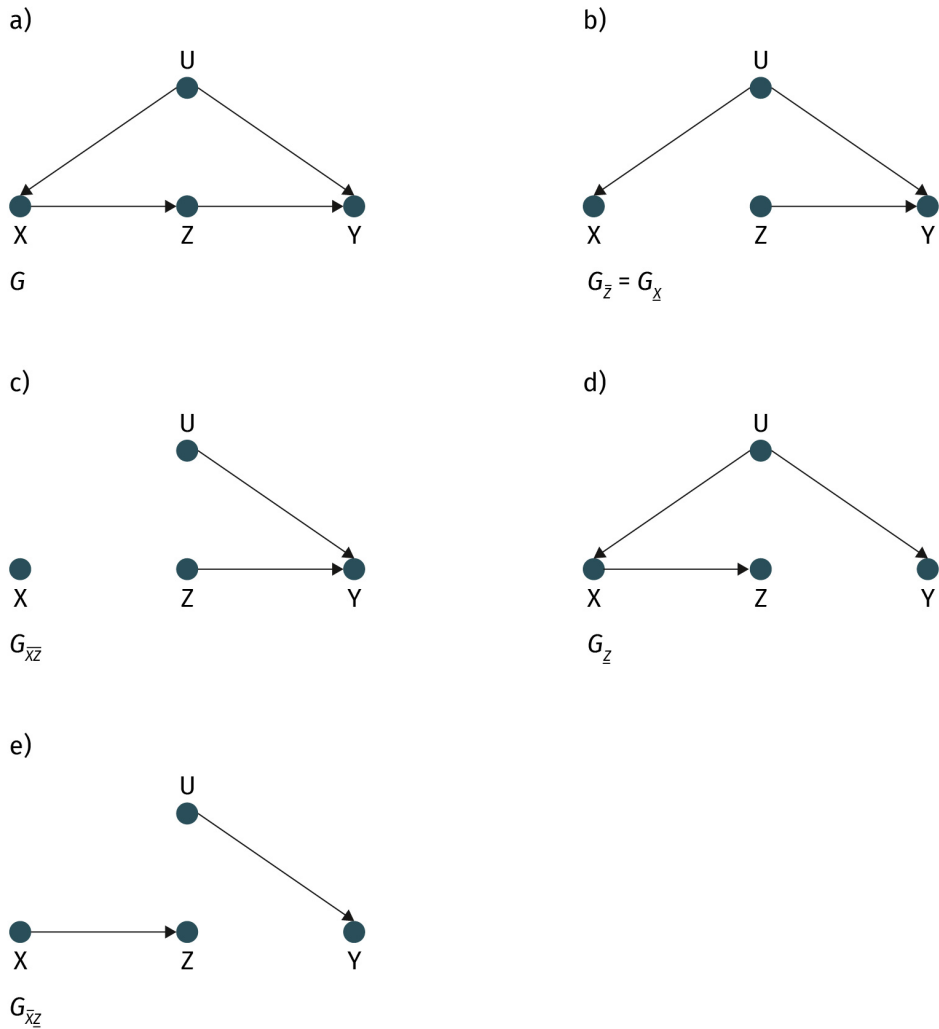
We also remember the definition of conditional independence we have encountered earlier: we let X, Y, Z be variables and $P(\,\cdot\,)$ a probability distribution.

The (sets of) variables X and Z are conditionally independent given Z if (Pearl, 2009, p. 11):

$$P(x\,|\,y, z) = P(x\,|\,z) \quad \text{whenever} \quad P(y, z) > 0 \tag{4.5}$$

which can be expressed using the notation $(X \perp\!\!\!\perp Y \mid Z)$. Informally, this means that once we know that Z has a specific value, learning the value of Y does not provide any further information about X.

📖 **THE THREE RULES OF** *DO*-**CALCULUS**

Let G be a directed acyclic graph that is associated with a causal model. For any disjoint subsets of variables X, Y, Z, and W, the following rules apply:

- Rule 1 (insertion / deletion of observations):

$$P(y|do(x), z, w) = P(y|do(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}}} \qquad (4.6)$$

- Rule 2 (exchange of action and observation):

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \quad \text{if}$$
$$(Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}\underline{Z}}} \qquad (4.7)$$

- Rule 3 (insertion / deletion of actions):

$$P(y|do(x), do(z), w) = P(y|do(x), w) \quad \text{if}$$
$$(Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}, \overline{Z(W)}}} \qquad (4.8)$$

where $Z(W)$ is a set of nodes Z that are not ancestors of any nodes W in $G_{\overline{X}}$.

(Pearl, 2009, p. 85)

The proofs for the rules of do-calculus can be found in (Pearl, 1995). The rules are a bit terse, so we will examine them in more detail. For more information, see (Pearl & Mackenzie, 2018, p. 234).

Rule 1 allows us to add or remove observations from our data. If we have observed some variable Z that is irrelevant (possible conditional on some other variables W) to the outcome Y we are interested in, then the probability distribution of Y will not change regardless of the value of Z— and the conditional probability for Y is the same with or without Z. That means the node for W blocks all paths from Z to Y. As an example, we can consider the fire alarm again. Since they do not detect fire directly, but via the presence of smoke, smoke is the mediator in the chain Fire $\longrightarrow$ Smoke $\longrightarrow$ Alarm. Once we know that there is smoke, we know the alarm will go off—whether or not there is a fire.

Rule 2 expresses that $do(X)$ is the same as $see(X)$ once we have controlled for all possible confounders. Informally, once we have removed all spurious correlations and closed all back-door paths, the remaining association we see in the data is the causal effect.

Rule 3 means that if there is no causal path with only forward directing arrows from a variable Z to the outcome Y we are interested in, we can remove the do-operation entirely. In other words, if we want to $do(Z)$ but it does not affect the outcome Y, the probability distribution of Y will not change, i.e., we will not cause an effect.

Following these rules repeatedly and in an appropriate order, we can express our interventions (symbolized by the do-operator) into expressions that can be estimated from observational data—if such a sequence exists, i.e., if the graph is identifiable. The good news is that these rules are complete and mathematically proven (Pearl, 1995). The bad news is that, while we can use the rules to verify that the sequence used to eliminate the do-operator is correct, it does not help us find the correct sequence, although algorithms exist for this purpose (Bareinboim & Pearl, 2012; Tian & Pearl, 2002; Shpitser & Pearl, 2006).

To show how the rules work explicitly, we return to the example of smoking we have solved earlier with the front-door criterion (Pearl & Mackenzie, 2018, p. 236): we wanted to determine whether smoking caused cancer in the presence of an unmeasured variable, the "smoking gene." The corresponding graph is shown in figure below where X corresponds to smoking, Y to cancer, and U to the unmeasured confounder (the smoking gene).

Figure 31: Example for Do-Calculus: Smoking
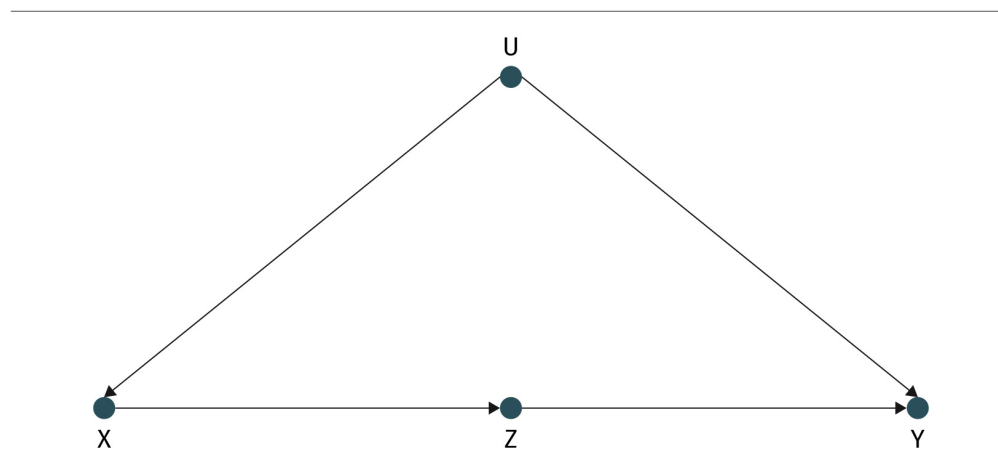
Because U is not measured, we cannot condition on it and hence we cannot block the back-door path, implying that we cannot use the back-door criterion and adjustment formula. However, as we have discussed earlier, if we include a new measurable variable (Z) (tar deposits) on the causal path from X (smoking) to Y (cancer), we can nevertheless establish the causal effect via the front-door criterion by conditioning on Z in the chain $X \longrightarrow Z \longrightarrow Y$.

Before we start with the example, we remind ourselves of the following relationship for the probability of an event (A):

$$P(A) = \sum_i P(A \mid B_i) P(B_i) \tag{4.9}$$

This means that the (total) probability of observing event A can be split into a sum of many conditional probabilities for events B_i, multiplying the conditional probability of observing A given that we observe B_i ($P(A \mid B_i)$) with the probability that B_i occurs, etc. This way, we can decompose the total probability of A into its dependencies of other events B_i that may be easier to obtain.

Let's now return to the example of smoking. We want to establish that smoking causes cancer, i.e., $P(Y \mid do(X))$. How would the probability of developing cancer change if we made the intervention $do(X)$, i.e., "make" people smoke. We do not want to do this in a random controlled trial—it would be unethical to force people to smoke and look who develops cancer with time.

First, we introduce the mediator (Z) using Eqn. (4.9):

$$P(Y \mid do(X)) = \sum_Z P(Y \mid do(X), Z) P(Z \mid do(X))$$

We now apply the second rule, which allows us to exchange intervention and observation if all back-door paths are closed. We remember that there is no back-door path between X (smoking) and Z (tar deposits), hence "seeing" is the same as "doing," and we can replace Z with $do(Z)$:

$$\dots = \sum_Z P(Y \mid do(X), do(Z)) P(Z \mid do(X)) \qquad \text{Rule 2}$$

There is a back-door path from Y (cancer) to Z (tar deposits) via the unobserved variable (smoking gene), but X (smoking) is a non-collider, and we can block the path by controlling for X. Hence, we can apply the second rule again and replace $do(X)$ with X in the second part of the sum:

$$\dots = \sum_Z P(Y \mid do(X), do(Z)) P(Z \mid X) \qquad \text{Rule 2}$$

Since we have introduced the tar deposits as mediator Z, there is no longer a causal path from X (smoking) to cancer (Y) once we intervene and "force" the tar deposits ($do(Z)$). Informally, we could say that, once we force tar deposits into the lung of the test subjects, it no longer matters whether or not they also smoke. Hence, using the third rule we are allowed to remove $do(X)$ from the equation:

| $\ldots = \sum_Z P(Y \mid do(Z))P(Z \mid X)$	Rule 3
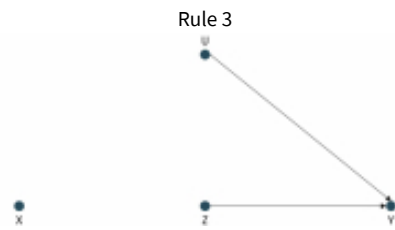
We now use Eqn. (4.9) again to account for all possible cases of smoking we control for, e.g., smokers and non-smokers or different amounts of tobacco consumed per day.

| $\ldots = \sum_{X'} \sum_Z P(Y \mid do(Z), X')P(X' \mid do(Z))P(Z \mid X)$	Eqn. (4.9)

Now we can use the second rule again, keeping in mind that the back-door path between Z (tar deposits) and Y (cancer) is blocked as we control for X (smoking). Hence, "seeing" is the same as "doing"

$\ldots = \sum_{X'} \sum_Z P(Y \mid Z, X')P(X' \mid do(Z))P(Z \mid X)$	Rule 2
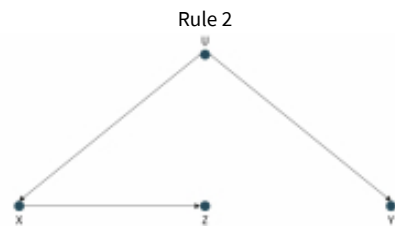
In the final step, we use third rule to replace $P(X' \mid do(Z))$ with $P(X')$. There is no causal influence from Z (tar deposits) to X (smoking), hence we arrive at:

$\ldots = \sum_{X'} \sum_Z P(Y \mid Z, X')P(X')P(Z \mid X)$	Rule 3
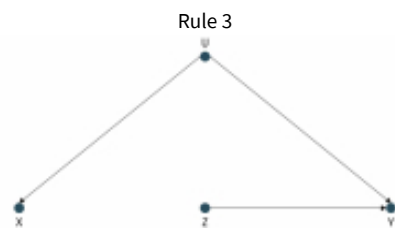
Putting it all together, we arrive at the front-door adjustment formula

$$P(Y \mid do(X)) = \sum_{X'} \sum_Z P(Y \mid Z, X')P(X')P(Z \mid X) \qquad (4.10)$$

where we have successfully replaced all expressions containing the do-operator into those without. These can then be estimated using the observational data.

As we can see, once we know the correct sequence of steps and rules to apply in each case, we can convince ourselves that the transformations are sound and follow the rules of do-calculus. However, as with most mathematical proofs, it will be quite hard to come up with the right sequence of steps.

📖 SUMMARY

The front-door and back-door criterion formalize the way we determine whether we can establish the causal effect of some variable X on an outcome Y. Tracing the back-door paths, we can determine if there are any spurious correlations between variables expected in the data and if we can block the paths to estimate the causal effects.

The back-door adjustment formula allows us to determine the causal effect from observational variables if the back-door paths can be closed. In some situations, the causal effect can be established via the front-door path even if the back-door paths cannot be closed due to unobserved confounders. In this case the front-door adjustment formula can be used.

The do-calculus formally expresses the mathematical operations that are required to transform expressions that contain the do-operator into those expressions that do not. However, in terms of the associated graph G we require that G is identifiable, i.e., if the causal effect can be established from observational data. The three rules of do-calculus are complete in the sense that they are sufficient to do this transformation if it is possible; however, the correct sequence and order of the operations is often difficult to ascertain.

HERE YOU CAN FIND THE VIDEOS FOR THIS UNIT.

Source: http://www.kaltura.com/tiny/g0stu

Alternative text:

UNIT 5

FALLACIES

5. FALLACIES

Introduction/Case study

Understanding complex systems is challenging and establishing causal relationships even more so. Exploiting correlations in the data can lead to very powerful prediction models that allow us to classify events or forecast future behavior. Indeed, the aim of machine and deep learning approaches is to exploit such correlations in the data to make accurate predictions. However, as we have discussed so far, correlations can be spurious, and variables can become associated because either we not not taken confounders into account or failed to block relevant back-door paths—assuming that we have already determined that there might be additional confounders or that we have created a causal graph for the task we wish to model.

In the following section, we want to highlight a few specific paradoxes, biases, and fallacies to highlight potential traps we wish to avoid when understanding complex systems. We also include a discussion about the imputation of missing data. This is a staple in statistical analyses, but common approaches used there typically do not take causal implications into account, potentially leading to different or even wrong conclusions compared to a causal model.

5.1 Mediation Fallacy

As we have seen earlier, mediators allow us to specifically express the way an effect comes about.

In general, we are mainly interested in the effect of an **intervention** (X) on an outcome or effect Y, i.e., $X \longrightarrow Y$. Mostly, we do not want to include an mediator (M) in the chain $X \longrightarrow M \longrightarrow Y$, as we are typically interested in the total effect of X on Y, for example, to see if smoking causes cancer.

However, in some cases, we may want to include a mediator, for example, if we cannot close the back-door paths because potential confounders are unobserved. In some situations, we may use the mediator to enable use of the front-door criterion. As we have seen in the smoking example, we could establish the causal effect via smoking $\longrightarrow$ tar deposits $\longrightarrow$ cancer.

If we include a mechanism that mediates the effect, we need to be sure that it is the right mechanism. This can be illustrated through the history of scurvy (Lewis, n.d.; Ceglowski, 2010) (Pearl & Mackenzie, 2018, p. 302ff): This disease was a major issue for early sailors on long distance trips across the Atlantic Ocean. It was found that a diet of citrus fruits prevented the disease. However, the way citrus fruit had a positive effect was never firmly established, and scientists assumed it was due to their acidity, i.e., citrus fruit $\longrightarrow$ acidity $\longrightarrow$ scurvy. However, a polar expedition undertaken much later was also compromised

due to participants contracting scurvy, causing much consternation. It was only later that the mechanism was discovered that prevented scurvy: vitamin C. Hence, the correct causal path is citrus fruit $\longrightarrow$ vitamin C $\longrightarrow$ scurvy. Adding a specific but wrong mediator leads to wrong conclusions.

There is, however, another reason to include a mediator in a causal chain:

we add a mediator if we are interested in the specific mechanism of an intervention. For example, let's say that we are testing a new medicine (X) and we want to establish its effect on the outcome of the patients (Y) based a specific way the medicine interacts with our bodies. In general, we can then split the effect of the intervention or treatment X on outcome Y into two parts, as shown in part a of figure below:

Figure 32: Controlled Direct Effect
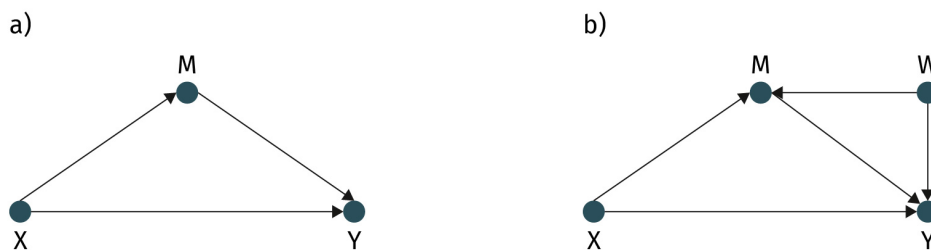
One path from X to Y is between the nodes directly, i.e., $X \longrightarrow Y$, and the other one comes via the mediator M, i.e., $X \longrightarrow M \longrightarrow Y$. In this picture, we could establish the part of the effect $X \longrightarrow Y$ by conditioning on M. However, this does not work in more complex graphs, for example, if there is a **common cause** of the mediator (M) and the outcome (Y), i.e., a confounder W. In this case, conditioning on M will block the path $X \longrightarrow M \longrightarrow Y$ and open the spurious path via the confounder: $X \longrightarrow M \longleftarrow W \longrightarrow Y$. Hence, if we do not condition on M, we cannot distinguish between the paths, including the mediator (or not). If we do condition on M, we condition on the collider along the path, including W, and introduce a new spurious association.

Common cause
A common cause of two variables is a parent to both.

There is no way to deal with this situation in classical statistics. However, the do-operator allows us to define a new concept of holding a variable constant without conditioning on it. Informally, we can say that we can obtain the "direct effect of X on Y when we 'wiggle' X without allowing M to change" (Pearl & Mackenzie, 2018, p. 317).

> **MEDIATOR FALLACY**
> The mediator fallacy occurs when conditioning on a mediator instead of holding the mediator constant. (Pearl & Mackenzie, 2018, p. 315)

The fallacy reveals that we intend to remove the influence of a mediator in establishing an effect from X to Y, but, by conditioning on it (rather than holding it constant), we introduce spurious associations between variables and confounders in the data.

Instead, we need to look at the controlled direct effect (CDE), where we intervene on the mediator M and force assign it to a specific value m and then compare the outcome where we intervene on the treatment $X = x$ or $X = x'$.

CONTROLLED DIRECT EFFECT

$CDE = P(Y = y | do(X = x), do(M = m)) - P(Y = y | do(X = x'), do(M = m))$

(Pearl et al., 2016, p. 77)

Note that the controlled direct effect depends on the value of the mediator M. For example, if all variables are binary and can take either 0 or 1 as its values, we can define $CDE(0)$ for the case where $M = 0$

$$CDE(0) = P(Y = 1 | do(X = 1), do(M = 0)) - P(Y = 1 | do(X = 0), do(M = 0)) \tag{5.1}$$

and, correspondingly, $CDE(1)$ for $M = 1$

$$CDE(1) = P(Y = 1 | do(X = 1), do(M = 1)) - P(Y = 1 | do(X = 0), do(M = 1)) \tag{5.2}$$

The expression for the CDE contains two do-operators. In order to estimate the controlled direct effect from observational data, these need to be removed. This can be done according to the rules of do-calculus. Taking part b of figure above as an example, we can do this by the steps listed below (Pearl et al., 2016, p. 77): There is no back-door path between X and Y.

Hence, since we control for X by comparing $X = x$ and $X = x'$, "seeing" is the same as "doing" (following the second rule of do-calculus). Therefore, we can remove the do-operator and the CDE becomes:

$$P(Y = y | X = x, do(M = m)) - P(Y = y | X = x', do(M = m)) \tag{5.3}$$

Next, we need to remove the do-operator on the mediator (M). Looking at the causal graph, there are two back-door paths from M to Y, one through the treatment X and one via the additional confounder W. The first path is already blocked, as we condition on X. The second path can be blocked if we condition on the confounder (W) (provided it is observable) according to the back-door adjustment formula. This results in the following:

$$\sum_{w} [P(Y = y | X = x, M = m, W = w) - P(Y = y | X = x', M = m, W = w)] P(W = w) \qquad (5.4)$$

and the resulting expression is free of do-operators.

Generally speaking, we can estimate the CDE of X on Y via M from observational data, i.e., the CDE is "identifiable" if the following conditions hold (Pearl et al., 2016, p. 77):

- There is a set S_1 of variables that block all back-door paths from mediator M to outcome Y.
- There is a set S_2 of variables that block all back-door paths from treatment X to outcome Y after deleting all arrows into mediator M.

We can also define the natural direct effect (NDE) using counterfactuals.

NATURAL DIRECT EFFECT (NDE)

$$NDE = P\Big(Y_{M=M_0} = y \Big| do(X = x)\Big) - P\Big(Y_{M=M_0} = y \Big| do(X = x')\Big)$$

(Pearl & Mackenzie, 2018, p. 318)

In the example in (Pearl & Mackenzie, 2018, p. 318), the authors use discrete binary variables, i.e. the variable Y takes only the value $Y = 1$ and the variable X takes the values $X = 0$ and $X = 1$. Informally, we can interpret the NDE as the expected change in Y when we change $X = x$ to $X = x'$ and keep the the mediators constant at the values they would have had under $do(X)$ (Pearl, 2009, p. 131). Additionally, we can define the Natural Indirect Effect (NIE) as the value when we hold X constant and set the mediator to the counterfactual value it would have had if we had changed X from x to x':

NATURAL INDIRECT EFFECT (NIE)

$$NIE = P\Big(Y_{M=M_1} = y \Big| do(X = x)\Big) - P\Big(Y_{M=M_0} = y \Big| do(X = x)\Big)$$

(Pearl & Mackenzie, 2018, p. 318)

Note that there are now two counterfactuals in the definition of the natural indirect effect, $M = M_0$ and $M = M_1$, whereas X only has the value $X = x$ (which is set to $X = 0$ in the example shown in (Pearl & Mackenzie, 2018, p. 318)), To explain the difference between the controlled direct effect and the natural direct effect, we follow the example of the Berkeley admission paradox (Pearl & Mackenzie, 2018, p. 309ff) (Bickel, Hammel, & O'Connell,

1975; Fairley, 1977): In 1973, Eugene Hammel looked at the graduate admission rates at Berkeley and noticed that, across the university, 35 percent of all female applicants and 44 percent of all male applicants were accepted.

He wanted to avoid any gender discrimination and, since graduate admissions (unlike undergraduates) were handled independently by each department, he looked at the values per department. However, once he did that, he found that women were consistently favoured over men, which seems paradoxical: How could overall admission indicate that men were favoured across the university but not when looking at each department that makes the decision? Looking at possible graphs, we start with part (a) of figure above, where X is the gender of the applicant, Y the admission to the graduate program, and M

the department. There are two paths from the gender to the outcome, one via the mediator (department) and one direct connection. As we have seen before, if this is indeed the correct causal graph, conditioning on M gives the correct results. However, this changes if

there is an additional confounder that influences both the mediator and the outcome. In this case, conditioning on M means conditioning on a collider, which introduces a spurious association.

If we were to look at the controlled direct effect (CDE), we would use the do-operator both on X and the mediator M, i.e., we would intervene on the gender and on the department.

However, if we truly did that, we would be forcing applicants to apply, say, to the physics department ($do(M)$) when they otherwise never would have. This would look very strange to the committee looking at these applications. Imagine an undergraduate student with a degree in, say, musical history applying to the physics department—they would most certainly not be admitted, as they lack the relevant previous studies. Instead, we look at the natural direct effect, where we let the students apply to the department they would have applied to anyway and then intervene on the gender. This is what is meant when we use the counterfactual notation $P\big(Y_{M=M_0}=y\big|do(X=x)\big)$. We look at the outcome, e.g., admission ($Y=1$), when the students choose the department ($M=M_0$) and then intervene to make them report for the purposes of this scenario, as one of two potential options—"biologically male" or "biologically female" as their sex ($do(X=1)$ or $do(X=0)$).

If the mediators are unconfounded, the natural direct and indirect effect can be estimated via the following adjustment formulae:
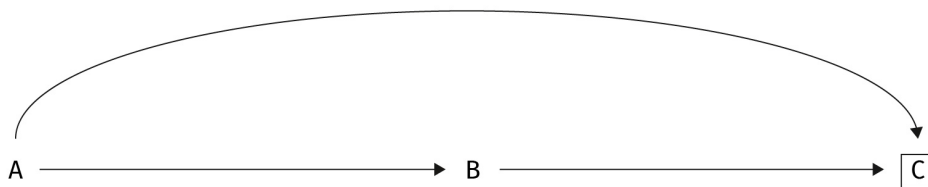
These adjustment formulae do not contain any counterfactuals or do-operators and can be estimated by looking at observational data.

5.2 Collider Bias

Collider bias occurs when we condition on a **collider** as shown in figure below:

Figure 33: Collider Bias

Collider
A collider is a node with two or more arrows pointing into it.

Both A and B are a common cause to C, and C is a collider, as arrows from both A and B point into C. The conditioning is indicated by a box drawn around C. We have already encountered an example of collider bias using the example of Hollywood actors (Elwert & Winship, 2014), where we used the graph: talent $\longrightarrow$ celebrity $\longleftarrow$ beauty. We can repre-

sent this using figure above when A is "talent," B is "beauty," we remove the arrow from A to B, and C is celebrity. For the general population, talent and beauty are unrelated. However, if we condition on C and only look at those who are celebrities in Hollywood, we find that the variables "talent" and "beauty" become associated. Intuitively, this can be explained in the following way: We know that the person is a celebrity. If their success is not due to talent, this makes it more likely that it is due to their beauty. Conditioning on a collider opens a previously closed back-door path between variables, which means they may become associated in the data. Note that this can also happen if we condition on descendants of variables that enter a collider, as shown in figure below (Pearl & Mackenzie, 2018, p. 160):

Figure 34: Collider Bias with Descendants
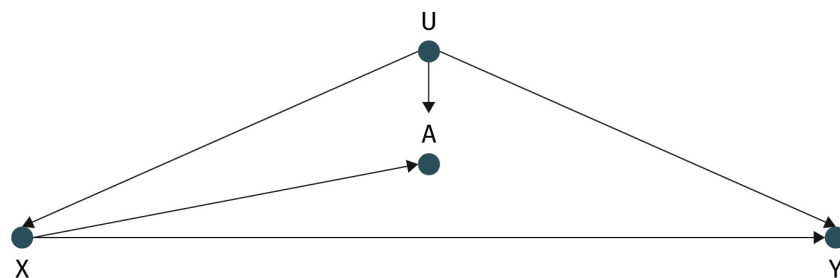
In this graph, U is a confounder of treatment X and outcome Y. If we want to establish the causal effect, we need to condition on the confounder. However, if U is unobservable, we cannot close the back-door path, meaning we cannot disentangle the causal effect. Since conditioning on a descendant of U would also close the back-door path (at least partially), we might be tempted to condition on A as a descendant of U, if A is observable. However, there is also an arrow pointing from $X \longrightarrow A$. Since arrows point into A both from U and X, A is a collider and conditioning on it would introduce a new collider bias and spurious correlation, even if it (partially) closes the back-door path of the confounder U.

Collider bias also occurs in the type of diagram shown in figure below which is called "M-bias" due to the shape of the graph (Pearl & Mackenzie, 2018, p. 161).

Figure 35: M Bias
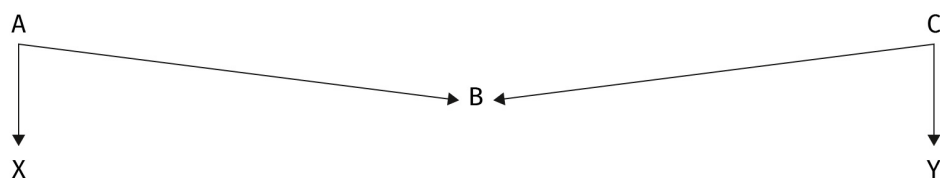
The variables X and Y are connected via a back-door path: $X \longleftarrow A \longrightarrow B \longleftarrow C \longrightarrow Y$. However, the path is already blocked by the collider B. Regardless, one might be tempted to call B a confounder because it is associated both with the treatment X (via A) and the outcome Y (via C). Additionally, it is not on a causal path from X to Y nor is it a descendant of an element of a causal path, because the graph does not have a causal path. Therefore, all three conditions of the test for confounders often used in statistics are fulfilled. Yet, it would be disastrous to condition on B, as this would unblock the path (because B is a collider). If A or C are observable, we can condition on either of them to close the path again should we accidentally or deliberately condition on B.

Collider bias can also be a source of selection bias. Selection bias is an umbrella term for biases that originate from the procedure by which we include individuals into an analysis (Hernan & Robins, 2020, p. 99). For example, a medicine (A) has a direct effect on the recovery of the individual (C) but the effect may also be mediated by a specific mechanism (B). If we only accept those into the study who have fully recovered from the illness, i.e., if we conditioned e.g., $C = 1$, we would unblock the collider and introduce a spurious correlation between A and B. We can avoid this by considering all individuals, regardless of whether they have recovered or not.

Another example of collider bias is the "birth weight paradox" (Pearl & Mackenzie, 2018, p. 183), (Hernandez-Diaz, Schisterman, & Hernan, 2006; VanderWeele, 2014). The data show that infants born in the United States whose parents (in particular, mothers) smoke are at a greater risk of lower birth weight and even death as compared to infants where the parents do not smoke. However, among infants with lower birth weights, the mortality rate is lower for those whose parents smoke as compared to those who do not. This sounds very paradoxical and counter to what we now know about smoking: If the parent smokes, their children have a better chance of survival compared to those of a non-smoker—if the infants have a low birth weight. However, if we draw the causal diagram shown in figure below, we understand that the apparent paradox is due to conditioning on a collider:

Figure 36: Birth-Weight Paradox
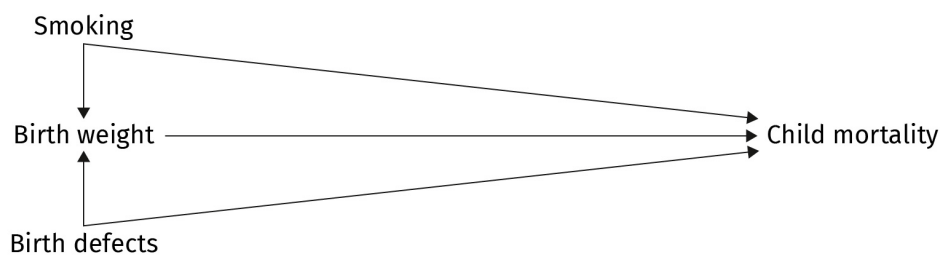
the parents' smoking both affects an infant's birth weight and increases their chance of death. Birth defects can also influence both the weight at birth and the mortality rate; additionally, the birth weight can also causally influence the mortality. Because smoking

can influence the birth weight and birth defects can also influence the weight, the variable "birth weight" becomes a collider. When conditioned on in the analysis, this introduces a spurious association.

5.3 Simpson's and Berkson's Paradoxes

When studying complex systems, we often encounter seemingly paradoxical behavior of variables that are associated with each other. Many of these examples are typical for a specific constellation in which we misinterpret the data and do not take the full (causal) story behind the often confusing behavior of the variables into account.

These paradoxes are often associated with a famous scientist who is associated with promoting or solving them in published works. In the following section, we will focus on well-known examples that illustrate how imperative it is that we analyze the data carefully and, specifically, think about the casual data-generating process.

5.3.1. Simpson's Paradox

Simpson's paradox is attributed to Edward Simpson, the statistician who popularized it. Essentially, the paradox describes a behavior seen in the data where a specific correlation between variables is observed when looking at the population from which the data are taken as a whole—but the correlation is reversed in every sub-population. This effect had already been observed by Pearson in 1899 (Aldrich, 1995) and later by others (Blyth, 1972; Cohen & Nagel, 1934).

We have already come across such an example earlier when we looked at the correlation between exercise and cholesterol, as shown in part a and b of figure below.

Figure 37: Simpson's Paradox
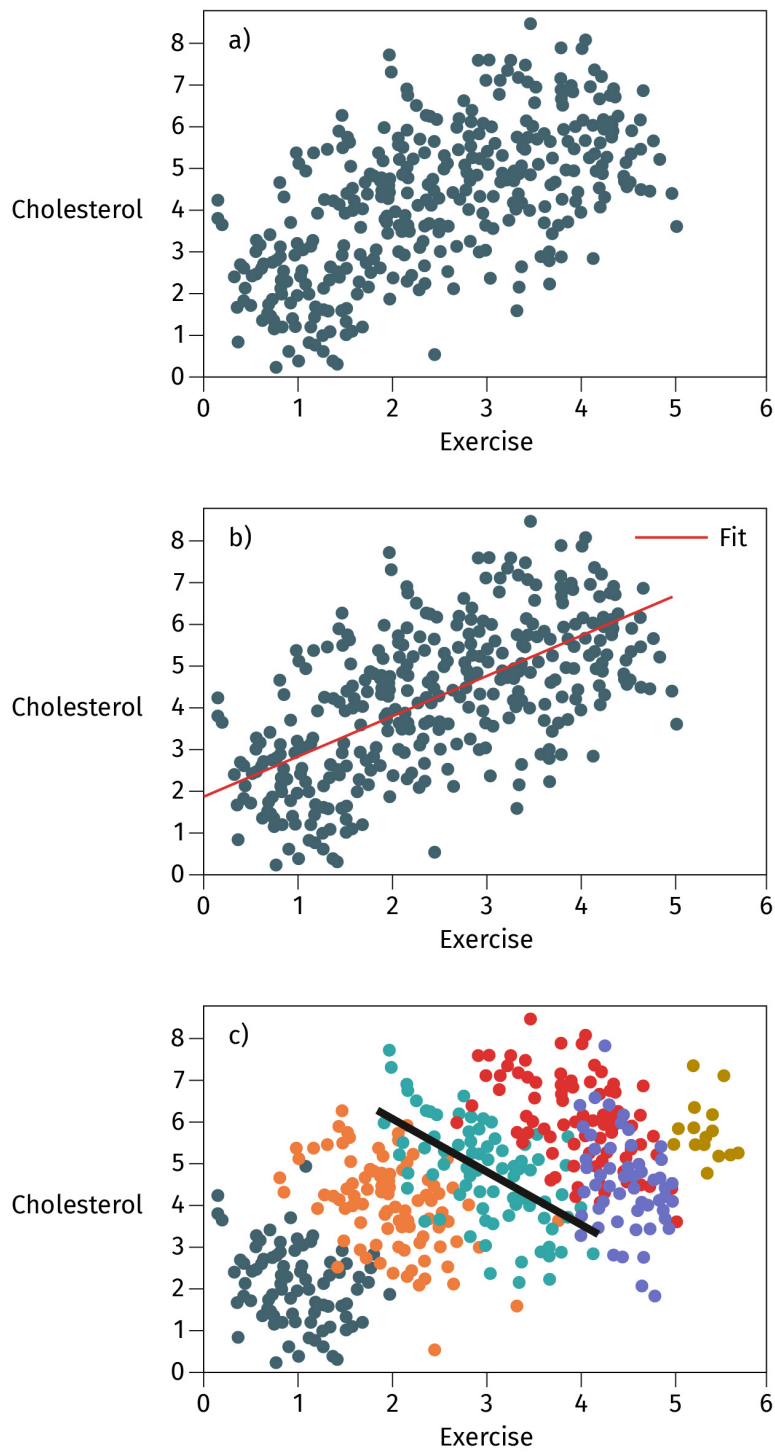
The two variables are strongly correlated, as indicated by the regression line. Yet, it is contrary to our general understanding that exercise is beneficial for us. We would expect that exercise, if it has any effect at all, helps to lower cholesterol, as it is generally beneficial to our health and high cholesterol levels are associated with health issues. Surely, exercising should not make it worse; however, this is what the correlation seems to suggest. However, once we look at the relation in different age groups, as shown in part c of the figure, the correlation is reversed. Instead, within each age group, we find the expected negative correlation between exercise and age.

The example given by Simpson who popularized the paradox is concerned with a new medicine that is administered to patients(Simpson, 1951). Note that the following example uses language written at a time that does not reflect today's standards concerning gender and sex. For its inclusion in this course book, we have kept the language as close as to the original as possible. Looking at all patients, fewer patients recovered who took the drug than those who did not. However, looking at the number of men, more men taking the drug recovered than those who did not, and the same holds for women. Hence, it seems that the drug helps men and women, but not if we do not know the gender. This is of course counter to any intuition we might have. To illustrate the example, we can use the following numbers (Pearl et al., 2016, p. 2):

	Medicine	No medicine
Men	81/87 recovered (93%)	234/270 recovered (87%)
Women	192/263 recovered (74%)	55/80 recovered (69%)
All patients	273/350 recovered (78%)	289/350 recovered (83%)

In this example, a total of 700 patients were enrolled in the study, 350 of which taking the medicine and 350 not. The first row seems to show that the medicine helps men: 93% of the men recover after taking the drug and 87% recovered who do not take it. The same is true for women: 74% recovered after taking the medicine compared to 69% who recovered without taking it. However, if we look at the data for all patients regardless of gender, only 78% recover if they take the medicine as opposed to 83% if they do not. This is, of course, paradoxical. If the medicine helps men and women, then, issues concerning gender or sex aside, it must help anyone.

We can write this as three statements:

- The medicine helps men and women.
- The medicine makes conditions worse for people.
- The medicine changes the gender of the patients.

Since the medicine most likely does not change the gender of the patients, one of the two other statements must be wrong.

The situation becomes a bit clearer if we draw a causal graph for the situation: We assert that the drug does not change the gender, but the gender may have an influence on the way the drug works. In fact, looking at the table above, we notice that the recovery rates are different for men and women. This also implies that we need an arrow from gender to recovery.

We also say that the medicine will have an effect on the recovery; hence we need an arrow from medicine to recovery. This is shown in figure below.

Figure 38: DAG for Simpson's Paradox (Confounder)
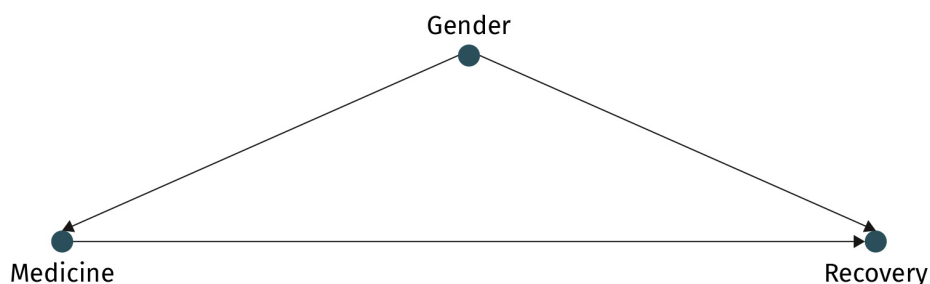
Now we can understand why the data seem to behave paradoxically: In this example, "gender" is a confounder, and we need to adjust for it to determine the causal effect and block the back-door path between taking the medicine and recovery. In the case of exercise and cholesterol, age is the confounder, and, if we control for age, we find that exercise is, indeed, good for our health.

However, we can also use the same numerical data in a different causal story (but with the column labels switched) (Pearl et al., 2016, p. 4):

|	No medicine	Medicine
Low BP	81/87 recovered (93%)	234/270 recovered (87%)
High BP	192/263 recovered (74%)	55/80 recovered (69%)
All patients	273/350 recovered (78%)	289/350 recovered (83%)

Now we assume we know that the medicine works by lowering the blood pressure—but it has a toxic side-effect. Now we see that the drug itself works. In the group that does not take the medicine, 87 patients have low blood pressure after taking the **placebo**, 263 have high blood pressure.

Among those who take the medicine, 270 have low blood pressure and 80 continue having high blood pressure. Hence, the drug does what it should: it "moves" the patients from high to low blood pressure. We also notice that the overall recovery rate of those who take the drug is better than those who do not: 83% compared to 78%. However, when we look

Placebo
A placebo looks and feels like real medicine but has no medical effect.

at the patients with low and high blood pressure, the correlation is reversed—but now blood pressure is a mediator rather than a confounder as in the case of gender earlier. We know the medicine is designed to work by lowering blood pressure, and we measure the blood pressure after the medicine has been taken. Hence, stratifying on the post-treatment blood-pressure disables one of the causal paths in which the medicine works, thus revealing the toxic side-effects. In this case, we should not condition on blood pressure, as it is the mediator of the effect (a chain in the DAG) and not a confounder (a fork in the DAG). Since we should not condition on "blood pressure," we look at the last line for all patients, which shows a higher recovery rate for those taking the medicine and would suggest one should take it.

Figure 39: DAG for Simpson's Paradox (Mediator) story
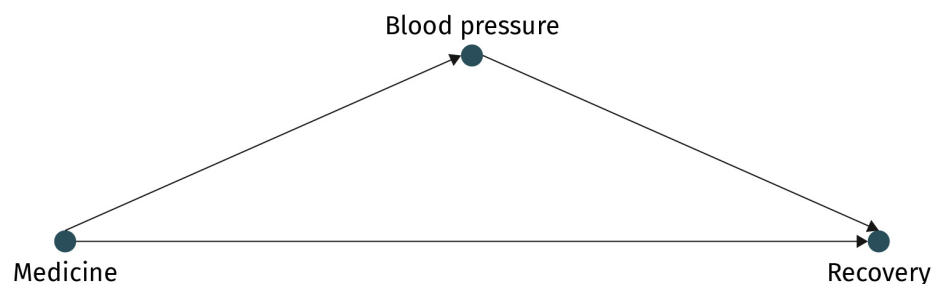
For further discussions about Simpson's paradox refer to (Pearl, 2014b).

5.3.2. Berkson's Paradox

Berkson, after whom this paradox is named, noticed an odd behavior of variables in observational studies conducted in hospitals (Berkson, 1946): Even if the occurrence of one disease is not related to the other in the general population, the two are correlated if we look amongst the patients in hospitals. The effect was studied over a long period of time and evidence collected, e.g., (Roberts, Spitzer, Delmore, & Sackett, 1978; Sackett, 1979)—however it was not clear why this correlation would come into existence.

To understand this bias, imagine we have only two diseases: disease 1 and disease 2. We can imagine two scenarios: In one scenario, having just one disease can be sufficiently severe as to require someone to be hospitalized.

In the other scenario, neither disease alone would require hospitalization.

Having both diseases, however, requires hospitalization. To illustrate how the paradox comes about, we draw the causal diagram shown in figure below:

Figure 40: Berkson's Paradox
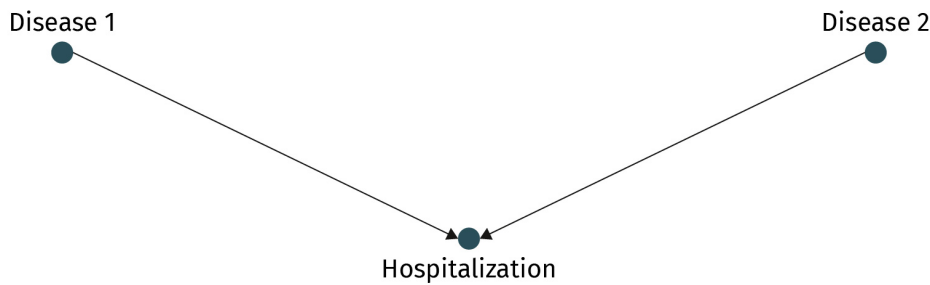
Disease 1 can cause hospitalization, hence we draw an arrow from "disease 1" to "hospitalization" and the same reasoning applies to disease 2. Disease 1 does not cause disease 2, and vice versa. Hence, there are no arrows between them. Looking at the causal graph, we can immediately explain why Berkson and others found a (spurious) correlation among hospitalized patients: By looking only at patients in the hospital, we condition on hospitalization ("hospitalization = true"). Because hospitalization is a collider, we open a previously blocked back-door path between "disease 1" and "disease 2." This can be also interpreted as a selection bias. Only those patients who made it into the study who were hospitalized; no individuals were randomly selected from the general population.

As mentioned above, we could interpret this collider or selection bias in two scenarios. In one scenario, either disease can be sufficiently severe to require admittance to hospital. This scenario is very similar to the example of celebrities encountered earlier: If the patient was not admitted due to disease 1, it is more likely they are admitted due to disease 2. In this case the correlation is negative. In the other scenario, hospitalization is required only if both diseases are contracted. In this scenario the correlation is positive: if a diagnosis confirms one disease, it is very likely the other is also present.

Further discussion is also found in (Pearl & Mackenzie, 2018, p. 197 ff.).

5.4 Imputing Missing Values: Causal versus Data-Driven View

Imputing missing values is a task that occurs frequently in the work of data scientists, statisticians, and all those who analyze data frequently.

Imputing missing data sees us determining a data point missing from the data we do have. This mostly this happens in **structured data**, which can be represented as a table. For example, a row in such a table may represent an observed event, and the columns would correspond to the variables we can measure that describe the event. In some cases, one or more of these variable values may be missing, and we need to account for this in our analysis. Naïvely, we might be tempted to just remove this observation with missing data.

Structured data
Structured data follow a defined structure called a "schema" and can be represented by a database table.

However, this could introduce a bias if the data are not missing due to a random glitch. There are a number of ways we can impute (or calculate) approximations to what we think the missing value should be—or at least determine a value that does no harm. For example, we can replace the missing value with the average value of all other values of the variable we observe in our data. We could also interpret the variable as a random variable and use all observed values to create an approximation of the underling probability distribution that governs the behavior of the variable. If the true distribution is known, we could fit its parameters from the observed data. Otherwise, we can create a non-parametric parametrization from the observed data to create an approximate probability distribution. This distribution can then be used to generate the missing value in a number of ways. For example, we can use the mean, mode, median, or any other quantile as an estimate—or we can draw a random number according to the probability distribution.

The latter approach has the benefit that it is not static, i.e., if several values are missing we use a different imputed one each time instead of the same value. These approaches are purely statistical, meaning we exploit no knowledge of the data generating process we might have. Instead, we only use the observed values only to infer the missing one.

Another option would be to find a matching pair of variables: Compare this event with the others, find the one that is closest to the one with the missing value for all other variables, and then use the matching event to fill the missing value. Note that, by doing this, we apply a type of conditioning because we require the value in the data to take certain values. We could also create a regression model to determine the values.

However, none of these approaches take the data generating process into account and include the causal story behind the data. Using the examples concerning salaries (Pearl & Mackenzie, 2018, p. 273 ff.) we will illustrate how the answers may be quite different if we follow a purely statistical or a causal approach to impute missing values.

The example uses the following table: Fictitious data for potential outcomes example

Fictitious data for potential outcomes example

| Employee (u)	$Ex(u)$	$Ed(u)$	$S_0(u)$	$S_1(u)$	$S_2(u)$
A	6	0	81,000	?	?
B	9	1	?	92,500	?
C	9	2	?	?	97,000
D	8	1	?	91,000	?
E	12	1	?	100,000	?
F	13	0	97,000	?	?

Here, Ex(u) represents the number of years of experience in a given job and Ed(u) the level of education where, for simplicity, it is assumed that only three levels exits: Ed=0: High school diploma; Ed=1 (undergraduate degree) and Ed=2 (graduate degree). S(u) represents the salary of the employee.

Note that this example is more advanced than the one discussed above concerning missing values, because here, we want to impute counterfactual missing data. Each individual employee has their own salary based on experience, as well as on their level of education. In this example, we want to know which salary an individual employee would have if they had a different level of education. Still, the same considerations apply. We could try to find matching entries for different employees and infer the counterfactual salary this way. Given sufficient data in all columns, we could also take any other statistical approach. For example, if we had a few hundred thousand employees, we could approximate the probability distribution of the salaries. We could also use a simple linear regression model for the salary: $S = 65,000 + 2,500 \cdot Ex + 5000 \cdot Ed$ (Pearl & Mackenzie, 2018, p. 274) in which each employee has a base salary of 65,000 that is then increased depending on experience (Ex) and the level of education (Ed).

However, as mentioned above, no statistical method takes the causal story or data generating process into account. They cannot, as statistical methods work with the data only. In a causal model, we would first think about the dependencies of the variables, i.e., which arrow points from one place to another. We can safely assume that both experience and the level of education have an impact on salary. For example, more years of experience and/or a higher level of education will, generally, lead to a higher salary. Therefore, we draw an arrow from education to salary and another from experience to salary. However, experience is also related to education.

Generally, we have two options: either an arrow from experience to education or from education to experience. It is more plausible that education is the cause of experience, i.e., that we draw the arrow from education to experience rather than the other way around. We could say that the level of education we have determines the years of experience we may have gained in our profession, whereas the number of years of experience will, generally, not affect our level of education. The resulting causal diagram is shown in figure below.

Figure 41: Causal DAG for the Salary Example
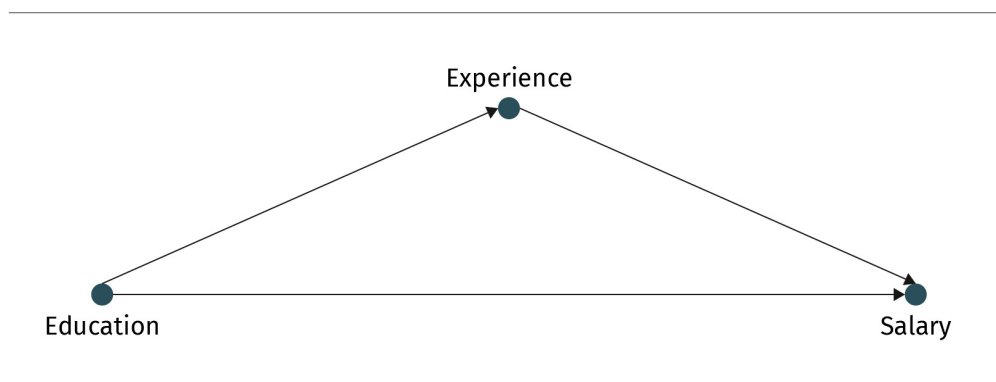
In this diagram, the years of experience are a mediator in the chain: education $\longrightarrow$ experience $\longrightarrow$ salary. However, if experience were a cause of education, experience would become a confounder, as the direction of the arrow would be reversed. This is important because we have to adjust for a confounder to avoid bias. However, we do not adjust for the mediator.

As we have done when discussing counterfactuals, we can translate the causal graph in a **structural causal model** (SCM) to calculate the counterfactual missing values. In such a model, the variable we want to model is a function of the related causal variables. In our case, the salary S is causally influenced both by experience and the level of education, i.e., $S = f(Ed, Ex, U_s)$, where U_s models any unobserved variations affecting the salary for a specific individual. In the simplest case, this can expressed as a linear model. In this example, we obtain (Pearl & Mackenzie, 2018, p. 277): $S = 5,000 + 2,500 \cdot Ex + 5000 \cdot Ed + U_S$.

Although the equations looks the same as the previous one (apart from the factor U_S), the interpretation is very different. Previously, we chose to regress S on Ed and Ex—but this had no connection to the real world. In particular, we did not assume a causal relationship between them. We could have chosen any other combination of the three variables. In contrast, our formula $S = f(Ed, Ex, U_s)$ now expressed our belief or knowledge that the salary (S) is causally connected to Ed and Ex. We cannot write a structural causal equation for, say, $Ed = f(S, Ex, U)$, because our causal model represented by the graph says that such a model does not exist. However, our model requires us to write another equation for experience: $Ex = f(Ed, U_{Ex})$, because we have added an arrow from "education" to "experience." The resulting equation is (Pearl & Mackenzie, 2018, p. 277) $Ex = 10 - 4 \cdot Ed + U_{Ex}$. Note that although we expect the salary and experience to be highly correlated in the data, the variable S does not occur in the above equation.

If we then want to know the counterfactual imputed missing value, we can follow the same approach we have taken when discussing counterfactuals.

Suppose we want to know the salary for employee A at varying levels of education. At present, employee A only has a high school diploma and six years of experience. How would it look if they had an undergraduate or graduate degree? In a first step, we use the structural equations to determine the unknown factors U_s and U_{Ex} for this specific employee and we find $U_S(A) = 1,000$ and $U_{Ex}(A) = -4$ (Pearl & Mackenzie, 2018, p. 278). Now we assume that employee had an undergraduate degree, i.e., we set the variable $Ed = 1$, or, in the language of causality $do(Ed = 1)$, and make the relevant change to the causal DAG by removing all variables pointing into Ed (in this example, however, there is no arrow to remove).

First, we evaluate the new level of experience for employee A using the second equation we obtained from the model $Ex_{Ed=1}(A) = 10 - 4 \cdot 1 - 4$, where we use the subscript to indicate that we calculate the counterfactual (i.e., hypothetical) for the case that we $do(Ed = 1)$, even though this is a hypothetical case. This means that employee A would only have two years of experience if they had an undergraduate degree. This can then be

used in the structural equation for the salary $S_{Ed=1}(A) = 65,000 + 2,500 \cdot 2 + 5,000 \cdot 1 + 1,000 = 76,000$. This is the salary employee would have if they had an undergraduate degree.

However, if we used the regression model discussed earlier, i.e., $S = 65,000 + 2,500 \cdot Ex + 5000 \cdot Ed$, we would get

$S - 65,000 + 2,500 \cdot 6 + 5000 \cdot 1 = 85,000.$

This is because we just changed the level of education and left everything else, in particular the years of experience, the same. Hence, in this example, we can get two answers for the imputed missing values, one from a purely data-driven approach leading to a salary of 85,000 and one taking the causal structure into account leading to 76,000. In the causal model, we also need to take into account the changed values of the variables we do not impute. In this example, we are looking at the years of experience the employee would have had if they had a different level of education, as well as the factors U that are unique to this individual. These factors are not considered in the regression model and, hence, the regression and the causal approach arrive at very different answers. This example uses a simple linear structural causal model, but the equations for $Ex = f(Ed, U_{Ex})$ and $S = f(Ed, Ex, U_s)$ may be more complex. This is especially true in more complex graphs with more than three variables.

SUMMARY

When we analyze data, we are often prone to make mistakes and fall into traps that we can only avoid by building a deeper understanding of the data and the data-generating process behind it. In several cases, the behavior of the data is seemingly paradoxical, for example, when the correlation between variables is observed to be one way across the entire sample but goes the other way in the subsamples.

Many of these paradoxes can be understood by analyzing the causal structure of the data-generating process. This allows us to identify why the data behave this way and how to avoid the paradoxical situation. Many of these phenomena are associated with the names of scientists whose relevant works analyzed the issue such as Simpson's or Berkson's paradox. When imputing missing values, we can take a purely data-driven approach or a causal approach. Using a concrete example, we can see how both approaches result in plausible answers, even though the results obtained by one approach can be very different to those calculated in the other.

HERE YOU CAN FIND THE VIDEOS FOR THIS UNIT.

Source: http://www.kaltura.com/tiny/iw9g6

Alternative text: